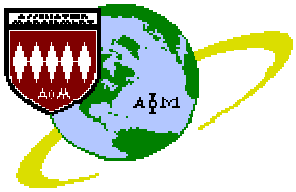

Journal of Information Technology Management

ISSN #1042-1319

A Publication of the Association of Management

ADVANCED APPLICATION OF BUSINESS INTELLIGENCE IN HIGHER EDUCATION: PREDICTIVE MODELING

DENNIS GUSTER
SAINT CLOUD STATE UNIVERSITY
dcguster@stcloudstate.edu

DAVID ROBINSON
SAINT CLOUD STATE UNIVERSITY
dhrobinson@stcloudstate.edu

CHRISTOPHER G. BROWN
SAINT CLOUD STATE UNIVERSITY
chrisb@stcloudstate.edu

ERICH P. RICE
SAINT CLOUD STATE UNIVERSITY
rier1201@stcloudstate.edu

ABSTRACT

The use of Business Intelligence or BI has become a hot topic amongst corporations looking to maximize profits by better understanding their customers. In this paper the authors look at unique ways in which the use of BI can aid institutions of higher education better prepare for incoming students, and develop ways in which to cut costs and improve the learning experience. To get the most out of the BI analytical process, it is necessary to involve and receive the backing of high level executives, without which the process will often languish. Once this necessary hurdle is crossed often the next major issue becomes how to structure the data repositories, through either a Kimball like Mini-Mart approach or in the view of the authors the more enterprise scalable Inman approach utilizing a Data Mart. Care should be given to whether existing data mining software solutions can be utilized, or if a solution created from the ground up is needed, based upon the institution's data and analytical needs.

Once the framework is decided upon, design and testing of the system can begin, including the important step of data cleansing. Without this necessary step the old adage "garbage-in, garbage-out" would undoubtedly take hold and the end-users who look to the system for useful BI will begin to lose faith in it. A Master Data Management or MDM would be the suggested method for assuring dynamic correction of data errors, although during the Extract Transfer Load (ETL) process this may prove to be too difficult if the data sources are of a highly disparate nature. Once the data has been placed in a useable format, the process of extracting useful information from statistical modeling can begin, and will require constant updating to assure the models stay relevant. If the entire process is well thought out and implemented, then BI can prove to be a highly valuable addition to any higher educational institution.

Keywords: Business intelligence, BI, Data Analytics, Predictive Analytics, Higher Education, Logistic Regression, Statistics, Data Warehouse, ETL

INTRODUCTION

The effectiveness of Business Intelligence (BI) as a support mechanism to support intelligent decision making is well accepted, Moss et al. [29]. As of late there has also been much interest in its application to higher education, Guster and Brown, [20]; Van Barneveld et al. [43]. In general the application of BI to any large organization should be a relatively straight forward procedure. However, in some organizations politics, differences in management styles and varying expectations sometimes can limit the effectiveness of BI, Lupu et al. [26]. The impact politics plays in higher education was observed by Guster and Brown, [20] and they found that it had a dramatic impact on many BI components from data dictionary definitions to the way data was staged. So therefore, one might expect implementing BI in higher education to be more challenging than in a private enterprise concern. In some cases the potential impact of deploying BI as a remedy for a poorly run organization within higher education has been unrealistic. There currently exists a balance between priorities and resource requirements for BI to be successful. BI initiatives need to produce results to continue to be a valuable priority for an organization. The challenges come when potential short cuts are presented as options that may create long-term resource costs and distract from building out an enterprise framework. An example of this is building out statistical models based on static data sets without initially investing in connecting the statistician's software to a Data Warehouse allowing provisioning dynamic data sources for modeling. Berg [9] points out that the predictive modeling techniques employed in BI are just tools not the definitive answer. Perhaps a good way to look at BI is to view it as a tool that can be used to make better informed higher education decisions. A good analogy to explain the value of BI might be the building of a house, you can build the house with hand tools, but you can build the same house much more efficiently with power tools and BI (and the use of analytics) of course would be analogous to using power tools, Pucher [32]. However, the success of devising effective analytical tools within a BI structure is dependent on several things. First, there must be an accurate and readily available data source. The old adage: garbage in garbage out is most appropriate here. Second, just implementing BI provides a very efficient structure for designing and implementing a computer system, but it is dependent on the underlying business logic. The wrong logic can be fatal. The work of Schonberg et al. [38] presents the essence of this concern, which is for BI to be successful the organization needs to determine what behavior indicates success. Further, that behavior must be quantifiable and recorded in such a form that it can be extracted in a timely manner to support the analytics strategy. If all of these conditions are met then BI can be effectively used to make intelligent decisions that will enhance the success rate.

To obtain an effective decision making strategy one has to realize that true analytics involves more than just generating a large number of reports. Rather it must focus on evaluating competing priorities such as maximizing student retention while shortening time to graduation, Goldstein [18]. In terms of what facet of educational management is appropriate for analytical analysis the literature states that it is appropriate in any facet related to ensuring operational success, Van Barneveld et al. [43]. Specifically, Natsu [30] stated that analytics could help educational leaders cut costs and improve teaching/learning. Further, she stated that the use of predictive analytics could range from improving efficiencies to saving money to enhancing student achievement. Some specific examples she cites include: planning courses, recruiting/retaining college students, optimizing the scheduling of classrooms and maximizing alumni donations. No matter the target management function for analytics to be successful the process needs to be understood and documented. In the early days of "data processing" a common technique to describe the flow of logic was flow-charting. While this technique has been revised and optimized the basic logic is still sound. Further, one often needs to coordinate the logic in regard to how the data is stored. Of course one expects the data to be stored in some type of database format and thereby one could categorize it as data mining. Therefore, it is important to select the data mining technique that will optimize the analytical formula being used. While there are several techniques, Laun [25], perhaps a Decision Tree is easiest to understand and illustrate the basic concept. Figure 1 below depicts a Decision Tree designed to document the decision process for examining Higher Education academic performance as it relates to High School GPA, Effective Financial Contribution (EFC), and ACT Range.

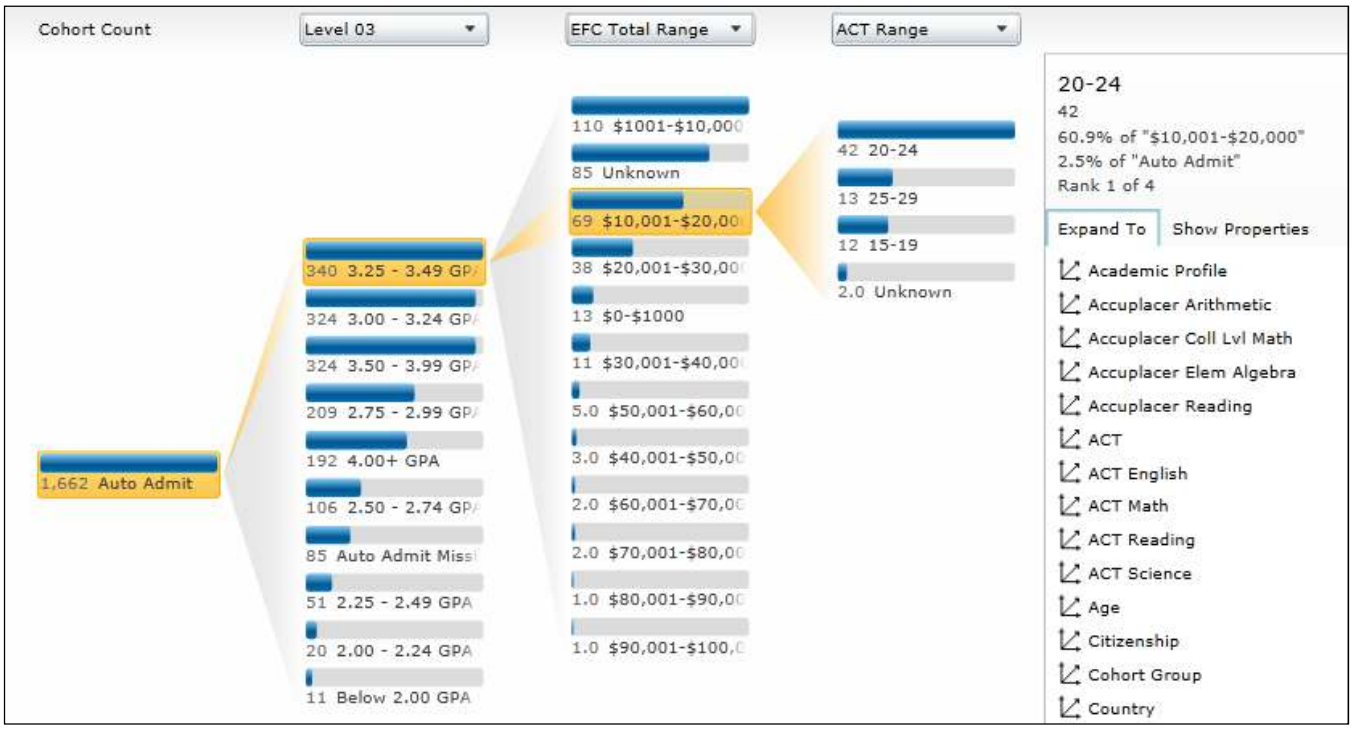

Figure 1: Academic Performance Decision Tree

Another critical point in devising an effective analytics strategy is to understand the limitations of the silo approach. In other words, the algorithms must be implemented on the enterprise level. Too often the models are devised, implemented and/or constructed on the individual, department or college level. According to Balkan and Goul [7], the silo-based approach doesn't take into account complex model inter-relationships, possible model correlation and covariance, as well as other independencies that can confound the resulting analysis.

Given that the analysis team is prepared to undertake complex data analysis on the enterprise level decisions need to be made about the modeling techniques. One of the first decisions is "how dependent will the team be on existing software packages"? There are certainly some viable options available that have proved effective. The work of Bao-sheng and Xin-quan [8] is a good example. They used Statistical Analysis System (SAS) Enterprise Miner to generate and operationalize predictive models for a Tele-communications Company. Starting with a software standard such as SAS Enterprise Miner has the advantages of having many well proven algorithms already pre-programmed while still having a framework from which to devise and implement your own algorithms. It is clear that a well-organized underlying database framework is crucial. Typically, the volume of data renders basic human visual analysis or traditional tools ineffective, Argotte, [5]. Therefore, specialized BI related algorithms and the ability to customize those algorithms are needed when devising successful BI analytics in any organization. In some cases BI software selection options are limited based on the degree to which the data source follows a conventional database design methodology and normal form. Often it is the proprietary design factor of the primary source data that precludes mainstream analytic vendor options, thus predisposing a homegrown solution. Another key factor to consider when selecting a build vs. buy decision relates to how well the business logic and definitions are understood by the analytics developers. Well defined definitions and ensuring business practices follow a standard definition that suggests a buy option may be optimal. However, if substantial data discovery is required to understand the business process and definitions, a build option may allow the flexibility required to satisfy current and future business requirements. Caution should be used when implementing a build solution to ensure designs are following

best practices to allow for sustainability and future vendor options as they become available.

Given the need to exploit the underlying database frame work the traditional approach to data analysis for decision support in which domain expertise is coupled with statistical modeling techniques to generate handcrafted solution needs to be expanded, Apte et al. [4]. Specifically, Apte et al recommend the concept of a Knowledge Discovery Database (KDD). This approach is designed to increase the availability of large volumes of data, facilitate the rapid deployment of data-driven analytics and deliver the results in a format easily understandable to end users.

It appears that Higher Education has accepted concepts such as KDD and tools such as SAS Data Miner. In fact Sahay and Mehta, [36], have devised an outline of options a University would be well advised to evaluate before implementing a BI analytics solution. Further, they address the issue of quality and introduce the concept of Quality Function Deployment (QFD) to help ensure the analytics will be effective across a wide variety of stakeholders.

The meshing of the underlying database infrastructure and devising high quality algorithms to evaluate that data is not a trivial task. Hopefully one starting from scratch would consult, Sahay and Mehta, [36], and use their step-by-step guidelines to get started. However, once the basic plan is devised, successful analytics require a long-term commitment in regard to fine-tuning the database infrastructure, as well as the analytical algorithms. The probability of getting it right the first time is next to non-existent. Further, Higher Education is in flux and constantly changing so the BI system must be designed as a dynamic system. Therefore, the purpose of this paper is to describe how the underlying database infrastructure was optimized to facilitate a series of data analytic algorithms. Further, the progression of those algorithms will be described and the algorithms delineated in detail. Last, an overview of the effectiveness of the resulting BI analytic system will be provided as well the interfaces used by end-users.

REVIEW OF LITERATURE

Key Players

Because BI is so complex, seldom do implementers devise all the code from scratch. So therefore, there is much reliance on prewritten software and hence the BI software vendor is a key player. Williams and Williams [44] report this assertion and state because of their vested financial interest that BI vendors are key players in the market expansion process through product innovation and their articulation of value propositions. They further state that key vendor initiatives include: (1) providing pre-integrated BI product offerings generally known as packaged analytical applications; (2) advocating expansion of the BI footprint within organizations, often referred to as BI for the masses and/ or enterprise analytics; and (3) positioning the use of their products as reflective of BI best practices. While the motivation for such initiative may lie in enhancing the vendor's bottom line, the degree to which a vendor is able to meet those initiatives can have a dramatic effect on how well the BI software functions and is cost effective. For example, software not written to support enterprise level computing and its massive amounts of widely distributed data would not meet the needs of a large company that uses enterprise level architecture. Further, because enterprise level computing encompasses so many elements to be successful it needs to take advantage of object oriented programming concepts and have all of the software tools integrated into a single package to provide maximum efficiency, Themistocleous and Irani [41]. Matching the software to the enterprise is crucial because today it is difficult to find a successful enterprise that has not leveraged BI technology for its business, Chaudhuri, Dayal, and Narasayya, [10].

Beyond the importance of software selection is the commitment of senior management Ramamurthy et al [34]. This involves support on the strategic level as well as appointing an effective project champion on the tactical level. It is also important to get the end-users involved early so that they understand the project and ultimately will support it Ang and Teo [3]. Perhaps the greatest influence on success is the composition and skill of the development team, Ramamurthy et al [34]. This typically might have the following composition: First, a Statistician would be needed to primarily function as a statistical modeler. This person would be responsible for the design, approach, and maintenance of the predictive models. Second, an Institutional Research Analyst would be needed to identify the data sources. This person would also be responsible for validating and creating data definitions from business logic. Third, Business Analysts would be needed to contact Data Owners and reconcile discrepancies in data errors or inconsistencies in business logic. Also this person would create Business Logic based on current business practices. Fourth, a Data Architect would be needed and responsible for operationalizing the Statistical model, data sets, analytic cubes, as well as the historical data. The Data Architect would also handle and notify others of data errors or schema violations as they occur. However, this could be better integrated into a Master Data management (MDM) and Data Governance Enterprise solution. Last are the key stake holders who would

help draw executive support, manage analytics resources and clear roadblocks at the administrative level.

Data Collection

Once the team of key players is in place and the appropriate software has been identified it is important to devise an effective strategy of data collection. The literature recognizes the importance of such effective strategies and Ramakrishnan et al. [33] have investigated the factors influencing such strategies. Specifically, they found that the strategies could be classified as either comprehensive or problem driven. They also categorized the BI purpose as either adding insight, consistency or to be transformational in nature. Of course, this illustrates a common problem with BI, which is a tradeoff between how quickly the system needs to come on line and how well it is planned out. Too often the goal is to get the BI system developed quickly, at the cost of being less cognizant of future applications and scalability limitations. Determining which data warehouse design methodology to follow depends on the commitment, priority, and the development team skill level. A Kimball [21] approach will deliver the quickest results; have the most universally adaptable application, most basic data architecture requirements, and provide sparse knowledge of the business process. However, the limitations occur when attempting to span multiple organizational units in a reporting solution. A well-conceived Inman [23] approach will have the greatest capacity for enterprise reporting with the greatest development cost and organizational commitment. Whatever the strategy it must focus on promoting a collaborative approach Andriole [2]. So once again the entire team needs to be involved in the process of identifying, storing, extracting and analyzing the data.

A first step might involve determining whether the data will come from disparate sources or be organized into a data mart structure. How the data is organized will have a dramatic effect in the success of the BI system, Ariyachandra and Watson [6]. It is clear that the data has to be accurate, consistent, complete, valid and timely, Cupoli, Devlin, Ng and Petschulat [12]. Further it is important to note that often a data warehouse exposes data quality issues inherent in the original source data systems. Hence the importance of data cleansing throughout the entire Extract Transform Load (ETL) processes cannot be over stated. If this cleansing does not take place before the data is made available to the end-users the BI system, then the data will be inaccurate and confidence in the system will wane. Once confidence is lost it is often very difficult to regain.

If data quality issues do appear during the ETL processing phase, Cupoli, Devlin, Ng and Petschulat [12] recommend the following steps be taken: define the data quality requirements, profile, analyze and assess the data quality, define data metrics, define the data quality business rules, clean and correct the data defects, and implement process improvements so that the data defects can be avoided in the future.

Given that the data will probably be coming from disparate sources it is important to be aware of any interdependencies within the data as they may ultimately influence data quality. Further, they may be useful in providing a quality check. For example, Guster and Brown [20] state that referential integrity events can be used as a validation mechanism before releasing the data from desperate sources to the data warehouse and dependent reports.

Before a data collection strategy can begin in earnest some serious thought needs to take place in regard to the refresh rate. Certainly it depends on the business. As an example, for e-commerce business it is critical that the data is as close to real time as possible to avoid inventory shortages resulting in the loss of valuable sales (and customer confidence). For higher education, the granularity can be larger. In fact, typically records need to be current only to the previous day. It is common to select a daily refresh strategy for the data warehouse. Execution MiH [15] state the reasons for this common refresh rate as:

- Integrating disparate sources conforming to differing reconciliation schedules.
- Avoid ETL processing during peak business usage.
- Common refresh rates are desirable to ensure uniform scheduling of updates to relationally constrained entities.

However, some business practices require more stringent refresh rates. For these instances it is important to recognize the impact of increasing the refresh rate and to ensure system performance (i.e. ETL cycle time) can meet the service level demand of the business practice. An example of scheduling data refresh cycles exists when processing employee expenditure reports. In the state record system certain operational tables are updated during Wednesday and Sunday night's batch processing. If dependent data is relationally tied to the expenditure data, proper synchronization is required to avoid a relational constraint violation. The time of day processing occurs can become problematic. The state operational database has a write database and a replicated database for reporting.

However, even with an indexed dedicated reporting database, running ETL during peak load can restrict business practices. Leveraging differential ETL processing could alleviate some of the processing perfor-

mance constraints. Determining the number of refresh cycles per day or per hour should be tied to the strategic objective associated with how the data will be used. An example of this is a student portal system featuring access to an electronic library reserves system based on currently registered courses. Requiring a student to wait until the following day to access course readings would have an adverse effect on student success and the repetition of the institution. Alternatively, data supporting budgetary planning reports may have a slower desired refresh rate of one week. This will allow adequate time to for review and revision. Another important concern addresses where the data should be housed both short and long term. This decision has major ramifications in regarding data extraction to accessibility and performance. Mehta, Gupta and Dayal [28] state that modern enterprise data warehouses have complex workloads that can be very difficult to manage. The trick is to devise a scheme to run these complex workloads 'optimally'. Further, Mehta, Gupta and Dayal [28], state that this problem has been evaluated from an Online Transaction Processing (OLTP) context. In this context, MPL (Multi Programming Level) is used as a means to reach an optimal solution. However, MPL can be difficult to implement in a BI scenario, since a low MPL can easily result in an under loaded system where as a high MPL can easily result in an overloaded system. Fortunately, the computing environments today offer flexibility in regard to where the data might be stored. It is possible to stage only the data needed for any given report outside of the primary data source. Duggan et al [14] recognize this trend and state that developments in data management systems, such as cloud and multi-tenant databases, are leading to data processing environments that can concurrently execute heterogeneous query workloads. However, this processing environment needs to be flexible to satisfy diverse performance expectations within the various applications. In meeting this need close attention needs to be paid to the expected Quality-of-Service (QoS) parameters and systems need to be continually watched and tuned. Again, thoughtful design architecture is paramount in providing a sailable system with performance capacity to meet the demands of the business logic while balancing QoS requirements.

Data Cleaning

The literature indicates that data cleaning is an integral part of the BI process. Florescuand [17] reports that there are three common types of problems: First, the absence of universal keys across different databases (the object identity problem), Second, the existence of keyboard errors in the data, and Third, the presence of inconsistencies in data coming from multiple sources. To combat these problems Florescuand [17] suggests a framework for data cleaning as a directed graph of data transformations. In this framework the transformations are placed in four classes: mapping, matching, clustering and merging. While the basic checking logic is implemented via something like a Structured Query Language (SQL) script the framework recognizes that human interaction is still needed on some level. Last Florescuand [17] considers the issue of performance which can be problematic as the size of the data increases. Specifically, the following optimization techniques were applied: mixed evaluation, neighborhood hash join, decision push-down and short-circuited computation. Further it is clear that, because the data will come from multiple sources and be used by multiple applications a data error not properly cleansed can affect the validity of analytical evaluation across the whole enterprise Scannapieco et al. [37]. The concept of a "data quality broker" is offered by Scannapieco et all. [37] as a way of keeping track of data stored on more than one source. This "data quality broker" would then evaluate the validity of each data source and ensure that the most accurate source was used.

It is generally accepted that an organization's data is a major asset. However, the commitment to test and cleanse data is typically less substantial. For example, a 2006 survey by Ambler [1] revealed the following:

- 95.7% thought the data was a corporate asset, but only 40.3% had a database test suite in place.
- 63.7% of respondents indicated that their organizations implemented mission-critical functionality in the database, but only 46% used regression tests to test the logic.
- 61.9% had production data problems, but 18% had no remediation strategy.

One therefore, might surmise that there is a disconnect somewhere in the decision making process. Once again the old adage garbage in – garbage out applies. Perhaps this can be attributed in part to upper level decision makers not understanding the BI process and thereby placing a disparate degree of resources in report generation, which is a tangible element they understand while short changing data warehousing and the cleansing process. Obviously this mistake can have dire outcomes.

Modeling

The nature of BI and the goal of taking all enterprise data and deeming it "big data" increases the importance of adopting a sound analytics strategy Chiang et al. [11]. The common understanding of big data is that it is unstructured whereby most of the warehousing frameworks are based on relational structures. This is also true,

and more inherently enforced, in Online Analytic Processing (OLAP) type models. Example: OLAP cubes require that every dimensional element in a fact table exist in the associated dimension table. When a referenced student ID (student unique identifier) does not exist the cube will fail to process unless ill-advised adjustments are performed. Further, it is important to zero in on the pertinent variables rather than using a "kitchen sink" method in which every potential variable is included Krupa [24]. The "kitchen sink" method tends to cloud the modeling process and adds noise when undertaking regression testing. However, over time as the modeling proves effective it is possible to add additional variables and evaluate or modify existing variables within the models.

The research indicates that statistical analytics has become an integral function within the BI process and the sophistication of those analytics continues to improve Segev et al. [39]. Further, it is important to address both individual and group trends as well as consider the granularity of the data Ferranti et al. [16]. In higher education BI this means that the characteristics of the individual students need to be evaluated as well as the characteristics of the cohorts (groups) to which they belong. Regardless of how well thought through and the analytic effectiveness, it will be critical to devise a user-friendly delivery system such as a dashboard so that end-users can take full advantage of them Neubock et al. [31].

Building and Testing

No matter how much care is exercised in modeling, the models need to be deployed and evaluated. The importance if testing models and assessing their predictive power is discussed by Shmueli and Koppius [40]. They further state that as models are built their purpose needs to be assessed in regard to whether they're intended to foster explanatory or predictive analytics. According to Graf et al. [19] there are needs in learning related analytics for both types of models. It is crucial that the model building and testing process be as objective as possible. However, too often there are external pressures that may limit this objectivity. Ramakrishnan et al. [33] state that once a model is proposed it should be tested based upon its appropriateness to the theoretical goal of the institution. Further, research needs to consider competitive pressure, as well as how well it will fit in a BI framework. Specifically, Ramakrishnan et al. [33] offers two data collection strategies (comprehensive and problem driven) and three BI purposes (insight, consistency, and transformation). Their schema provides a theoretical lens from which to enhance and understand the motivators as well as the factors related to collecting and analyzing the "Big Data" needed to be successful in implementing BI.

Operationalizing

One of the key factors in operationalizing business intelligence is sound integration of BI and Business Processes (BPs). Marjanovic [27] states that this is important involving cases in which BI aims to unify strategic and tactical decision making, by integrating BI solutions with an organization's constantly evolving BPs. However, operationalizing BI is not always an easy undertaking. Specifically, Isik et al. [22] report that one of the reasons for failure is the lack of an understanding of the critical factors that define the success of BI applications, and that the BI capabilities within the implementing staff are among those critical factors. It is therefore crucial to operationalize what constitutes success and how to measure that success Ranjan [35].

Reporting and Publishing

While this often viewed as the last step in the BI process it may be the most important. However, the large-scale of many BI implementations can make this a difficult step Ulanov et al. [42]. Further, it is critical that the reports are in a format that the end-user can understand which means the implementing BI staff need to be flexible in meeting individual needs or all prior work was in vain. Ulanov et al. [42] call this attention to detail in matching the reports to the end-user "personalized reporting". Certainly there are numerous options for generating these reports, as there is even a wealth of open-source reporting tools. An excellent diagram that reports best practices of matching the tool to the end-user is provided by Denny [13].

DESIGN AND IMPLEMENTATION

Key Players

The authors are affiliated with a Midwest regional comprehensive state university system herein referred to as the university. The application described in this paper closely follows the roles identified in the literature review. The BI and Analytics work is distributed among five positions in addition to two or three part time student positions. It is often the case when a single position is responsible for several roles. The Associate Vice President & Associate Provost for the Office of Strategy, Planning & Effectiveness is responsible for prioritizing projects and ensuring effective strategic alignment within the organization. A Faculty Statistician with partial release time is responsible for developing the statistical model used to generate predictive and descriptive analytics. The Statistician also leads a team of student researchers contributing to the validation, data collection, and creating

the statistical model. Business logic is maintained by the Institutional Research Analyst through validating BI data with Institutional Research data sets and resolving business logic and or business process inconsistencies. In essence Institutional Research is acting as a Business Analyst. Database systems are administered by the Systems Administrator and Business Intelligence Engineer. Responsibilities of the BI Administrator include all database maintenance tasks as well as ensuring security models are effectively applied. The BI administrator spans multiple roles serving as a Business Analyst, DBA, Data and System Architect. The Business Intelligence Data Architect ensures historical data is tracked, relational and multidimensional models accurately address the business questions for self-service and reporting requirements. Ongoing commitment to addressing data inconsistencies and quality problems as they are identified continues to be a challenge. It is the authors' observation that validating data definitions and business logic along with data quality continues to be the single greatest time factor impeding operationalizing analytic projects. Definitions are challenged as they are tested against relational data. Business rules are often discovered during the operationalization phase requiring revalidation of the business rule and data conformity checks. Errors are brought before the business analyst to reconcile, which may impact the relational design and require schema changes. A single challenge or business logic contradiction can result in several days of corrective action. Attention to validation of business logic and reconciling data quality problems early in the process results in exponential gain in workflow efficiency. Data reporting and provisioning for self-service clients is a shared reasonability of the BI Administrator and BI Architect. Ideally a BI Developer would assume the role of report writer and self-service provisioning.

Data Collection

The authors struggled with the selection of a Kimball [23] versus and Inman [21] approach. When selecting a warehouse design, one needs to consider the reusability and maintenance of current and future BI initiatives. If implementing a Kimball approach one would expect to have many customizable templates for each project. An example is ETL processing. Each data source table is explicitly defined and pulled into a staging area. A second template is then customized for the transformation process. If a change is made to the data source, each dependent template must be adjusted to reflect such changes. Maintaining architectural consistency can be challenging as nomenclature and convention may change across organizational units. One common problem might include an unknown member in dimensional tables. Another approach might rely on the multidimensional engine to dynamically create the unknown dimensional label. Challenges arise when combing these two subtle design differences into an enterprise multidimensional project. An Inman approach can leverage modular design processes. Rather than customizable templates, an ETL engine can be implemented accepting parameters for each table. This can greatly reduce the data integration and maintenance time. Data source changes can be corrected in the ETL module rather than changing customizable templates associated with the data source. Process errors can also be addressed in a single location. Learning curves are reduced when self-service users are presented with enterprise data sources following consistent conventions and providing homogeneous user experiences. Enterprise definitions are inherently supported, enforced and maintained. The institution described in this paper chose to follow an Inman approach. The main deviation in the authors' implementation from the literature is observed through the challenges realized in the lack of an institutionally adapted data dictionary and a sound formal data governance policy. Much of the business logic was forged during the data mart design phase resulting in a much delayed development rate. Another concern relating to maintaining and gaining end-user confidence was realized through the inability of the development team to demonstrate progress through delivery of self-service data reporting solutions during the design and development phases. This became increasingly problematic as BI reporting waned in its ability to effectively assist with institutional decision making and dynamically responding to administrative data inquiries. Users would find alternative paths to collect data for reporting purposes, many of which did not conform to institutional definitions. This prevalence of data shopping resulted in discrepancies among reports. Often users would request the same report from several reporting agencies and select the one best supporting their desired outcome. As more reports are made available along with accompanying definitions the expectation is supporting report definitions will become the de facto standard for institutional data based decision making.

Data Cleaning

The system described in this paper underwent many data cleansing operational setbacks including: Untrusted source data; duplicity; business logic not validated against a relational schema; contradicting states (student concurrently accepted and denied); nonexistent dimensional reference (Previously awarded degree no longer exists in the source validation table); and lack of form validation during data entry. Much of the cleansing pro-

cess required customized detection, reporting and data isolation methods. Tested data cleansing processes were then transformed to an automated or operationalized implementation. At any point in the analytics process detection of poor data quality can result in retooling the relational schema, redesigning dimensional models, altering the business logic and associated reports. The degree to which preventive measures are taken to protect the data warehouse from erroneous data is correlated to the data governance and protective measures implemented in the Operational Data Store (ODS). Data sources with unenforced referential integrity are more likely to require considerable data protection procedures during ETL processing. Whereas, sources following Codd's third normal form are more likely to require less data quality processing and tend to be easily operationalized into a BI solution. It is not uncommon during the data validation process to uncover business process responsible for data errors. In this case, recommended corrective action may not be confined to the data and may also include making adjustments to business processes. Often reporting back the discrepancies or separating out the errant data for corrective action can often prove beneficial in keeping the ETL process running and avoiding schema validation errors. Best practice suggests a Master Data Management (MDM) solution, allowing for dynamic correction of detected errors. This states when errors are detected there needs to be a way of correcting the error in the source systems at time of detection. Having this escalated up to a manual process of manually editing the source operational data is not considered a viable MDM solution. Inherent in the definition of MDM is the requirement of correcting errors when they are detected. Without error correction in place it would be impossible to bring up a viable MDM solution let alone sustain one. This would require a formal data governance presence within the organization and adequate dynamic corrective action capabilities within the ODS. Security, staffing resources, and legacy ODS systems may preclude a full MDM solution.

Modeling

The authors' predictive analytics implementation plan was based upon three interdependent modeling aspects. First, a dimensional model is discovered through a working session with the stakeholders, Business Analysts, and the Data Architect. The Architect that determines which data elements are not currently available in the warehouse and either makes a recommendation to add data elements to the warehouse and schedule ETL processing, or recommends the analyst derive a one-off data set for the purposes of data exploration and discovering key statistically significant variables. Second, a Statistical model is fitted to the projected inquiry. This is dependent on the dimensional model and should align with the dimensional modeling process in identifying dependent variables and fitting the appropriate predictive model to the data set. The selection of candidates for statistical significance is not trivial. Selecting too many variables for a model or the wrong variables can result in noise and cloud the significance of otherwise potentially useful variables. Several elements are essential to the successful transition to the final modeling stage. The statistical model must be able to be operationalized. For example, if a regression equation is used to create a probability of each student returning on a subsequent term, then each of the included variables must be matched to dynamically available data within the warehouse. Any missing variables will have to be translated to database entities and added to the warehouse. In some cases this requires extensive business logic validation and testing to ensure data elements are appropriately aligned and any schema constraints are not violated. Additionally, the grain must be agreed upon. An example where this can create a reporting problem arises when projecting a count of the number of student that will be in the incoming fall freshmen cohort. If a statistical model is used based on applying deltas to the number of applicants at that time compared to previous cohorts it will be problematic when looking at a student level dimension or demographic and measuring the inferred probability of, for instance, students with a Gross Family income less than $45,000 to return. The best practice is to use one model at a higher cohort grain to validate the sum of probabilities at a student grain. The student grain probabilities are then integrated into the warehouse and used on the final modeling phase. From a multidimensional modeling perspective, student grain is preferred in order to leverage associated demographic and affinity relationships not inherent in the model. When the grain is at a cohort level, it is much harder to synthesize these relationships into the data mart with relevance for decision making purposes. Third, a multidimensional model is constructed to facilitate the analytics reporting and longitudinal projections across varying dimensional elements. This takes the form of cubes, mining models, or other analytic frameworks. It is common during this phase to make adjustments to the previous phases and potentially discover additional untapped statistically significant variables to be added to the data set and integrated into the statistical model. Throughout the entire process, data validation and cleansing techniques are continuously applied in order to mitigate the compounding effect of errant data propagating through the modeling phases. The Data Architect will often take into account the considerable overhead of repeated cleansing and validation into consideration when determining to request a

one-off data set or move toward integrating missing data elements into the data warehouse prior to the statistical modeling phase. The details of this implementation plan are illustrated in Figure 2.

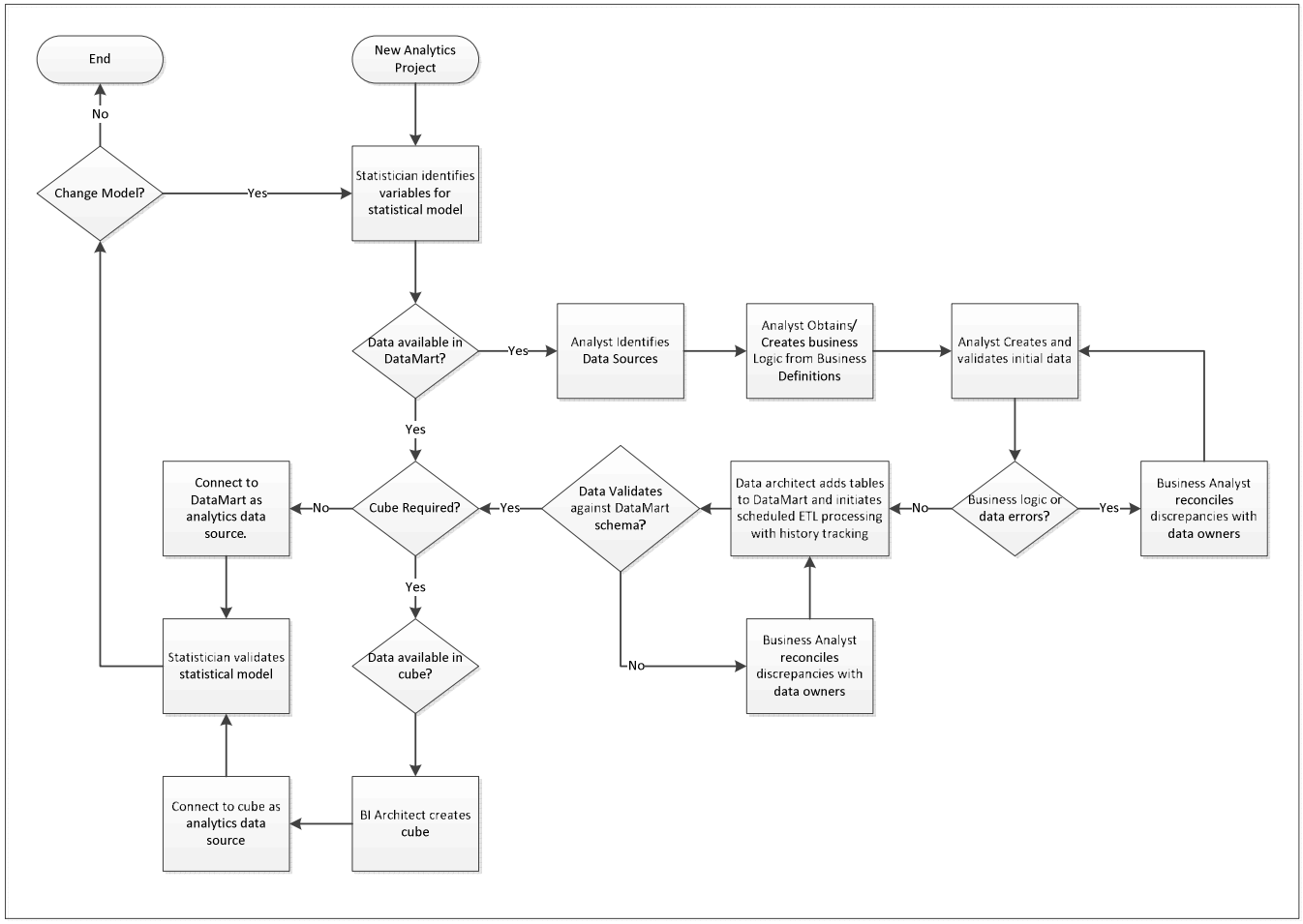

Figure 2: Analytics Modeling Process

Building and Testing

The modeling addressed in this paper closely follows the process identified in the literature review. Throughout the modeling process testing is performed on prospective result sets. Results are validated by the analyst against the original dimensional model to ensure they accurately represent the inquiry. In addition the dimensional model is tested to ensure it continues to strategically align with organizational business processes. An example of this occurred during a modeling exercise in attempt to examine the statistically significance of new entering freshmen high school GPA on third term retention. A range was applied to the high school GPA. Upon further progression in the modeling process, the initial ranges were not accurate enough to make business process changes regarding admittance of lower GPA students. The range was then expanded and a new data set was generated to apply more precision to the statistical variable. Once the result set was validated the business process was changed regarding admission criteria for lower performing applicants. The term bit-bucket is barrowed from the manufacturing industry whereby parts that do not pass initial inspection are set aside for closer examination or sent back to be reworked. In some cases, when irreconcilable errors are discovered in the operational data source, a bit-bucket is used to classify and partition out errant data. Typically this incurs a status of "Unknown"

for the purpose of multidimensional reporting. When Data owners or Business analysts are alerted to the number of dollars or students classified as Unknown, they are motivated to correct the source data, which in future ETL processing, persists to apply appropriate reclassification of the bit-bucket item into a more organizationally applicable category. This cyclical data quality improvement process provides continuous validation opportunities.

Operationalizing

The techniques and success criteria for determining operationalizing strategy closely follow the literature review in this paper. Two distinct approaches are taken to determine which course of action to take when operationalizing analytic models. Factors used to determine the appropriate approach include: Complexity of the model; fluctuating coefficients verses variable selection; number of disparate data sources; availability of data in the current warehouse environment, and if a requirement for historical longitudinal data exists. When dealing with a highly complex model containing many levels of dependent sub steps, it is valuable to map out a relational data structure to ensure data integrity is maintained during the build process. Often unnoticed contradictions in business logic are uncovered while fitting data to a relational schema. An example of this occurred when building cohort sets. New entering freshmen and new entering transfer students had to be unique. This meant a transfer student can only be considered a transfer student if not in a previous new entering freshmen cohort. Several occurrences were identified of transfer students in previous cohorts. Previously the test for unique cohort placement occurred at a per term basis, not across all terms. An institutional Data Mart can be used to pull relational data together from desperate sources for institutional reporting purposes. As smaller department level or mock version MiniMart can be used to rapidly build up a data structure. The relational constraints in the data schema allowed for a higher integrity in adhering to the defined business logic. It is important to consider these factors when determining to operationalize into a Data Mart model or a minimart design as stated by Guster and Brown [20].

When building a statistical model the selection of statistically significant variables is critical to the accuracy of the predictive results. Often variables are added or removed during the tuning process. For example, regarding a regression model, this variability can impact the coefficients in the model. Coefficients can also be impacted by changing data within the data sets. If the statistical model had high degrees of fluctuation in determining which variables to include in the model, then a minimart design may be appropriate until or if a more stable model is identified. If the fluctuation occurs in the coefficients with little or no variation in the selection of the variables then direct Data Mart integration would be more appropriate.

The number of disparate sources also impacts the decision to operationalize directly to the Data Mart or use a minimart model. If more than one data source is used building the ETL processing into a Data Mart is preferred. Also if the majority of the fields required are available in the Data Mart, then adding remaining fields to the Data Mart is preferred over recreating a minimart for the model. Longitudinal data requirements should also be considered when selecting which operational model to leverage. Longitudinal or historical data requirements including slowly changing dimensions are a good candidate for a Data Mart rather than a minimart model. Due to security and performance considerations, in some cases a hybrid approach is used populating a minimart from the Data Mart. An example of a case when this hybrid approach is preferred is when reporting Enrollment by course by day comparing term day over a five-year period.

It is the authors' recommendation, although requiring more upfront development and design work, directly operationalize into the Data Mart unless there is expectation the model will require significant tuning.

Reporting and Publishing

The following consists of the culminating step in the authors' implementation process. Typically reporting is considered the final step in the modeling process. If an agile or iterative approach is used the reporting process may occur as a validation of the process or a proof of concept. Some models will require revision and tuning to be operationalized. Pivot reports are typically used to validate the process as business logic is infused in the ETL processing during operationalization.

During the publishing process various levels of user interaction and access are considered. Some reports may be at an enterprise or institutional level accessible by the executive administration. Executive reports are typically provided in the form of a dashboard or business scorecard. In the higher Education examples provided in Figures 3, 4, 5 and 6. Dashboards are used to provide high-level overviews of Key Performance Indicators (KPI). Strategically identified KPI provide drill down capabilities to allow self-service users to persist the data to underlying supporting disaggregated forms. Other mid-level reports may be filtered by department for information targeted reporting. Other parameter driven reports may be extremely flexible providing a self-service window to the data. The reporting format and audience should be considered at all levels of the modeling process

to ensure the architecture can support the reporting framework from an accessibility and security context. Finally data quality or business logic anomalies require immediate reconciliation. Delays in addressing data quality factors may prove detrimental and require a significant magnitude of dedicated resources to correcting and rebuilding data models, schemas, business processes, and multidimensional reporting structures.

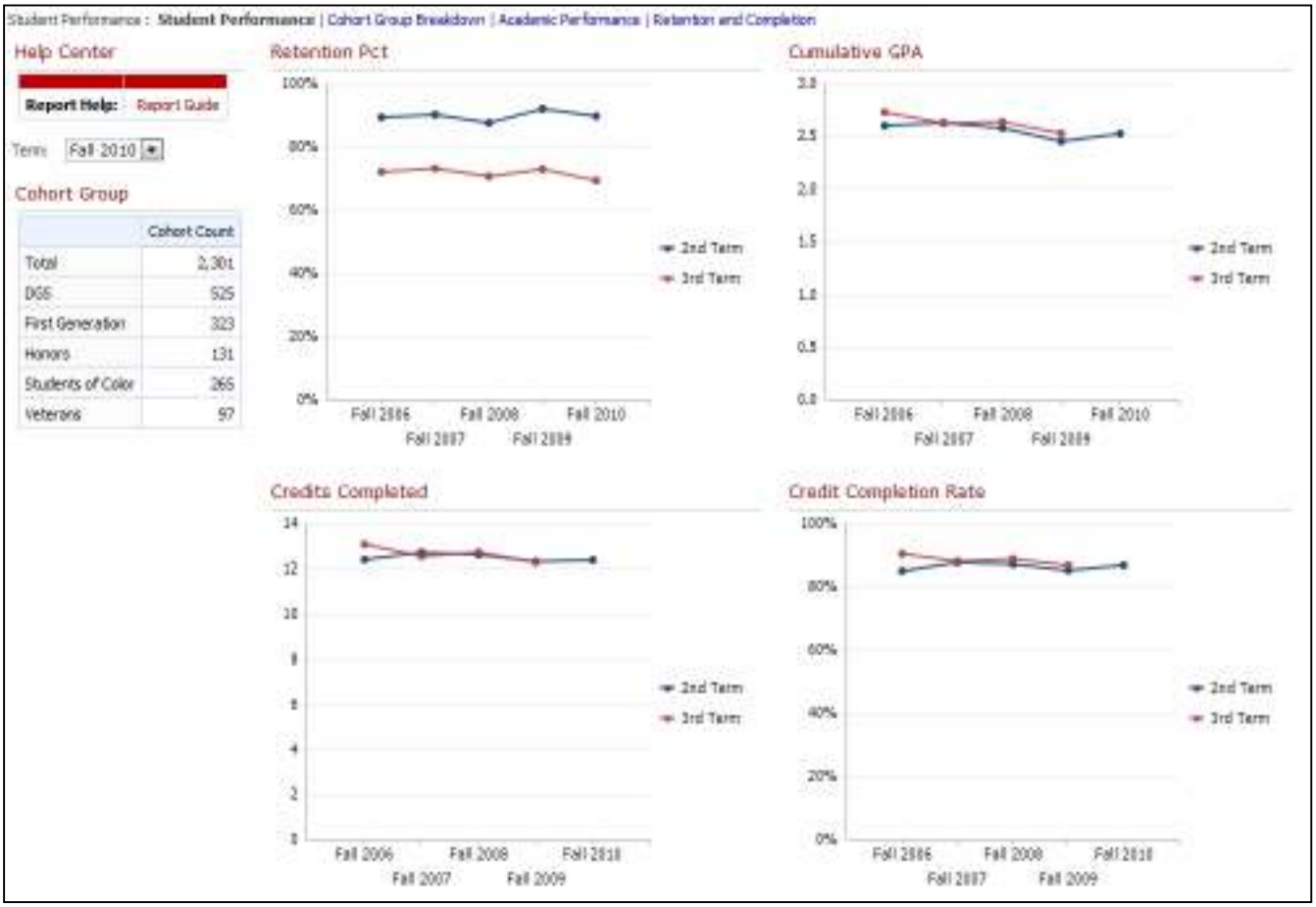

Figure 3: Student Performance Dashboard

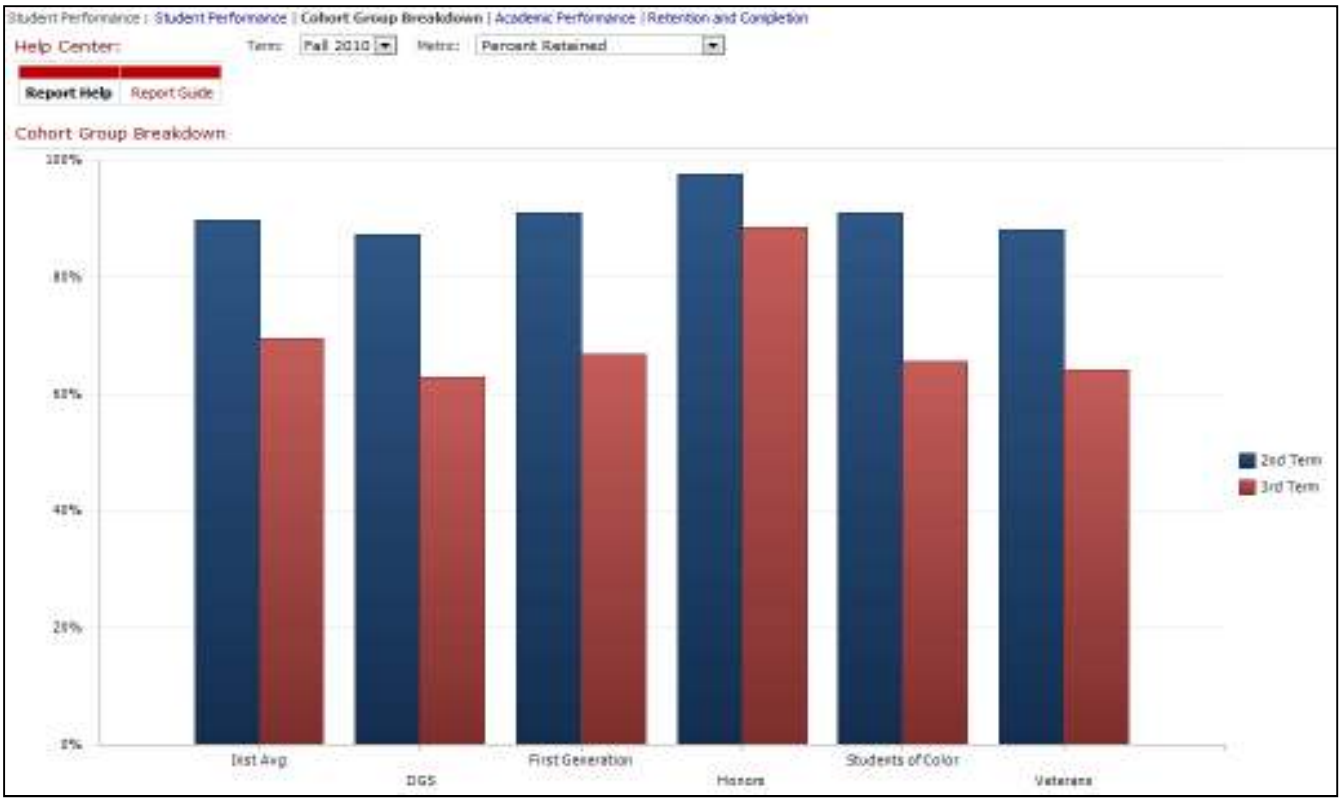

Figure 4: Student Retention Cohort Breakdown

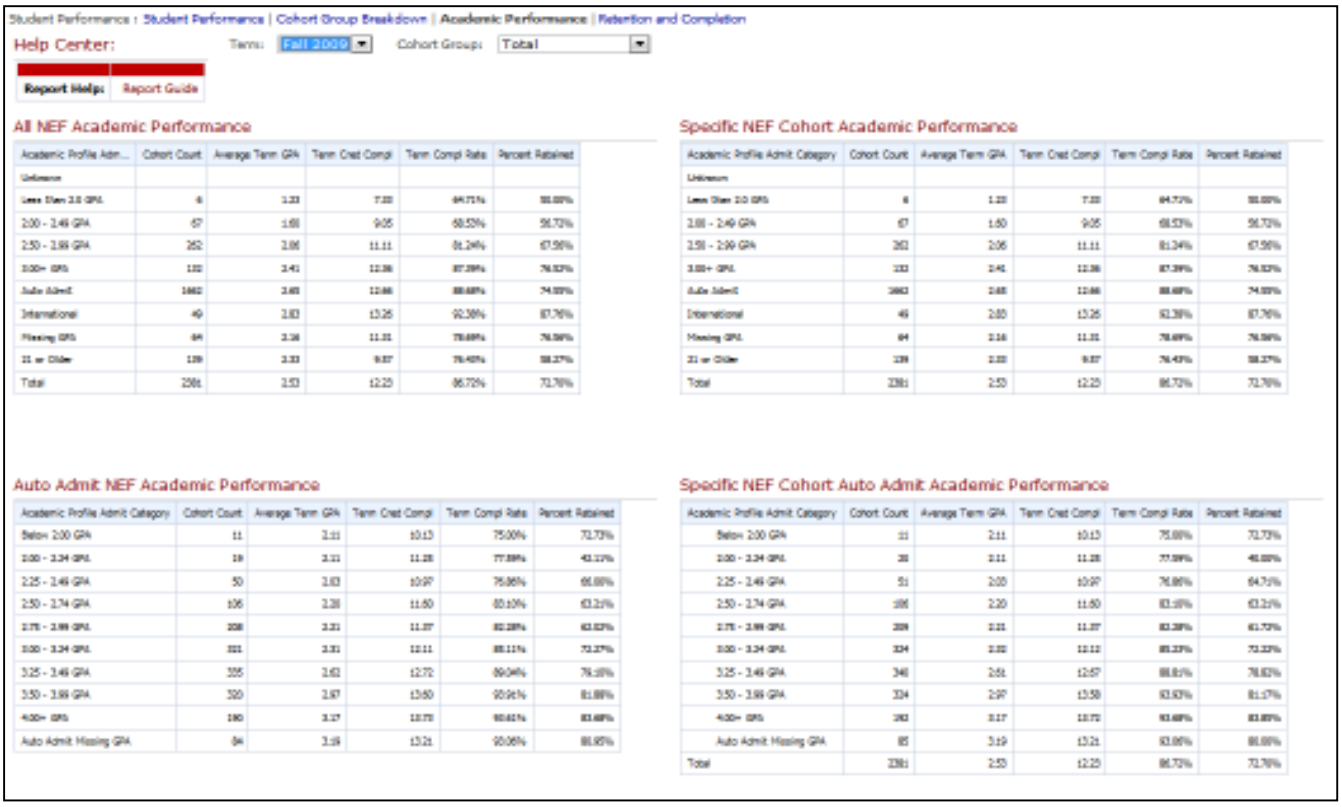

Figure 5: Student Academic Performance

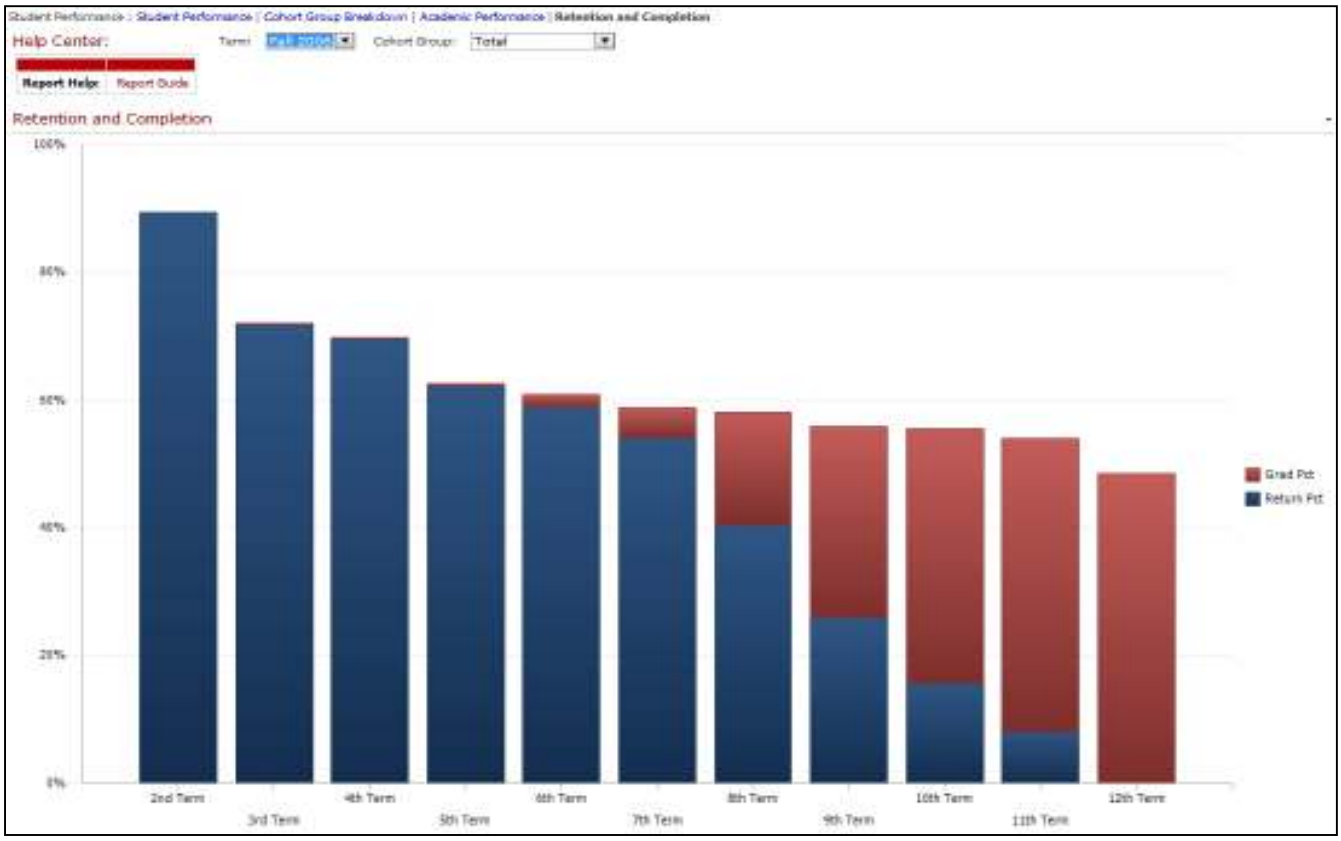

Figure 6: Student Retention and Completion

Due to the increase demand for data driven reporting solutions and the lack of available supporting data governance structures the reporting work described in this paper diverges from the best practices recommended in the literature review. A consultant was contracted to assist with designing a reporting framework for delivering strategically relevant interactive reporting. As the Data Mart was not fully operational, several hybrid Kimball models were rapidly developed to support the reporting framework with the expectation that the underlying model could be operationalized and the report and multidimensional sources could be redirected to the operationalized data source. The results were not anticipated. When the reports were presented to the user community several recommendations prompted the rebuilding of relational and multidimensional models. One example occurred when presenting intended and declared majors. The original reporting solution provided separate reports for intended and declared majors. The user community wanted the ability to combine the intended and declared majors to account for anticipated course seat requirements for the upcoming terms to allow ample time for requesting faculty lines. Having the intended and declared majors listed on separate reports was problematic as the naming convention for majors differs from intended majors resulting in unassociated intended majors to their major counterpart. A satellite operational database was developed to maintain the mapping of majors with intended majors. The relational schema was then changed by adding an intended major field to the Major table. This prompted a change in both the multidimensional model and supported reports. Unanticipated challenges arise when undergoing significant relational schema changes. One such challenge to consider is how to treat historical data. One might truncate historical tables to the data the schema changed. This have potential undesirable effects including impairing the ability to conduct longitudinal analysis. A more preservation focused approach may involve the costly process of creating a custom transformation procedure to convert the previous data schema to the current model. It is the authors' recommendation that historical transformation only be pursued when a strong business

case is made requiring persistence of historical data for archival purposes or longitudinal analysis. As a best practice under data governance, a formal data retention policy should be maintained.

DESIGN AND IMPLEMENTATION

To provide the reader with the details that go into devising and operationalizing analytic models this section describes some of the methodology used by the authors. Although the models provided are by no means inclusive of all the techniques that might be employed. Rather, they are designed to provide a representative sample of some of the problems that might be encountered by an institution of higher education that is attempting to leverage BI in their data analysis.

Data Cleaning and Normalization

Without reliable and accurate data the validity of any analytic modeling would be suspect. Unfortunately, data cleansing is one of the most time consuming tasks of the data analyst. Often it is not completed, because the lure of moving ahead to analyze or run models is too great. However we try to resist that lure to the extent possible with our student data. There are cases with incorrect entries for certain fields. When detected, the incorrect value is replaced by a correct one. Sometimes it is not known what the correct entry should be; in those cases, the incorrect value is deleted and the field is empty. This has occurred when a student's birthdate is recorded to be after their graduation date, or when a student's high school GPA is outside the range of acceptable values.

Some student data is not consistent with others because of differences among high schools. In this case, the data must be normalized, that is converted to a common scale or set of categories. Grade point average (GPA) is a prime example. Most high schools grade on a 4.00 scale, meaning that the highest GPA possible for students from those schools is 4.00. A lot of schools now give more than 4 quality points for an A in advanced class. So in about 1% of the applications received at the university, the student actually has a GPA higher than 4.00, even though 4.00 is nominally the maximum. A few high schools have a scale for grades that is different from a 4.00 scale. One scale that occurs with moderate frequency is a 12 point scale. So a student may have, for example, a GPA of 10.50 on a 12 point scale. In order to use this kind of information as a variable in a statistical model, it is important that the data represent consistent measures across all cases, even though the high schools have diverse means of calculating grade point averages for their students.

Our solution to the problem of the 12 point scale is to rescale the values down to a 4 point scale. In general, if a high school reports grades on a scale with a maximum of M, then we compute a rescaled High School GPA, which we call "HS GPA 4 Scale." It is computed according to:

$$\text{GPA 4 Scale} = \left(\frac{4}{M}\right) * (\text{HS GPA})$$

where HS GPA is the grade point average reported by the student's high school.

Dealing with GPAs that are reported as larger than 4.00, even though the high school has only a 4.00 scale, is a problem requiring an ad hoc solution. We have two possible solutions:

1. Leave the GPA as it is, for use in modeling.
2. Change the GPA to a new "Adjusted GPA" with a value of 4.00, in these cases.

There are certain advantages to either of these solutions. By leaving the GPA unchanged, the information inherent in the exact value is still present. For example, a student with a 4.50 GPA on a 4.00 scale is clearly an exceptional student. Probably the student has taken many Advanced Placement courses or done other kinds of accelerated college-level coursework. The 4.50 GPA reflects the true nature of the student's ability. The disadvantage is that not all high schools compute their GPAs to take into account this level of difficulty in a student's coursework. Thus similar students at other schools may have only a 4.00 GPA or even slightly below a 4.00 GPA. So changing all GPAs above the 4.00 maximum to be 4.00 would correct for much of this inequity between schools, but at the same time has the disadvantage of masking predictive information, which may be present in the higher GPAs.

Handling of Missing Data

Some of the variables that are collected from and used as part of our student database include missing values. For example, not every student files a (Free Application for Federal Student Aid) FAFSA report to determine their financial aid eligibility. Some students don't submit American College Testing (ACT) scores, because they can be admitted by qualifying in other ways. Some high schools don't rank their students, so that High School Rank and High School Percentage Rank are not available for some of the students. When using multiple regression or logistic regression models based on student data, any student's data that is not complete will be omitted from the analysis. Because there are several important variables in which might contain missing values, the number of students omitted from the analysis can be substantial. This tends to weaken the analysis by reducing the sample size

of available data. Furthermore, it is often true that the presence or absence of a particular value can be predictive of the response variable. For example, students filing FAFSA reports are more likely to enroll as a new student than students that haven't filed FAFSAs. It is critical that a system be in place to account for these students with missing data. While there are multiple means of doing this, we have found one particular method that has the flexibility to work in a variety of models and situations.

Define two new variables (often termed "dummy variables" in statistical jargon) for each predictor variable with missing values:

$DV1$ is defined to be 1 whenever the value of the predictor variable (e.g. ACT Composite score) is known, and 0 whenever it is not known.

$DV2$ is defined to be the value of the predictor variable (e.g. ACT) whenever it is known, and 0 whenever it is not known.

These two dummy variables are used in the predictive models in place of the original variable (ACT). In essence, this can have the effect of allowing separate models for students that have a known score versus students that have the score as missing. For example, in the case of a regression model using ACT, a usual regression model might be constructed as:

$$Y = B0 + B1 * ACT + error$$

Here the variable Y might represent some measure of student performance at the college level, for instance college GPA. Because all students with no ACT score would be omitted in the calculation of this model, which would be a significant portion of students among university applicants, a preferred method of using ACT as a predictor would be the following:

$$Y = B0 + B1 * DV1 + B2 * DV2 + error$$

Because of the method of defining $DV1$ and $DV2$, the model can be specified for the two groups of students.

Group 1: Students with known ACT

$$Y = (B0 + B1) + B2 * ACT + error$$

Group 2: Student without known ACT

$$Y = B0 + error$$

For Group 2, the predicted value of Y would be a constant, independent of ACT, while for Group 1 the predicted value would (typically) increase as ACT increases, in proportion to the value of $B2$.

Other methods of replacing missing values, often called imputation, exist. Some researchers use the following procedure:

$DV1$ is defined to be 1 whenever the value of the predictor variable is missing, and 0 whenever it is known. (Note that this definition is the opposite of that presented above.)

$DV2$ is defined to be the value of the predictor variable (e.g. ACT) whenever it is known, and the overall average of that variable whenever it is not known.

Group 1: Students with known ACT

$$Y = B0 + B2 * ACT + error$$

Group 2: Student without known ACT

$$Y = B0 + B1 + B2 * (Mean\ of\ ACT) + error$$

This method has the effect of replacing all missing values for ACT by the mean ACT score. This means that no imputed values of ACT are introduced outside its normal range. In some cases this is an advantage over the previously mentioned method, especially when the students are not missing values (in this case, ACT scores) for identifiable reasons, or the observations with missing scores are not thought to be atypical when compared to the other observations. By keeping the imputed values in the middle of the range of the ACT values, no outliers are artificially introduced into the data, which may result in better estimation within this region of the data space.

A third method of imputation is similar to the preceding one, but is simpler in that DV1 is not used; only $DV2$ is used. This results in Group 1 and Group 2 being estimated with similar models:

Group 1: Students with known ACT

$$Y = B0 + B1 * ACT + error$$

Group 2: Student without known ACT

$$Y = B0 + B1 * Mean\ of\ ACT + error$$

Here we assume that the students without ACT scores can be assumed to follow the same model as the students with ACT scores. This assumption may be reasonable in some cases, but is not reasonable in many situations. For instance, in our experience and at our institution, students missing ACT scores tend to be very different, atypical of the normal student. For the authors' university, this method would not be acceptable.

Not all of our predictive models are regression models of this sort; other useful predictive models include logistic regression models and classification tree models.

These models can also use dummy variables as defined above in a useful way, to represent data that is missing and data that is present. This use of information about what is known and not known is a powerful tool in the statistical analyst's set of methods.

First Iteration of Predicting Enrollment of New Entering Freshmen (NEF)

Ratio Estimation of NEF

The tools that are used for predicting student enrollment are among the most useful predictive methods in operation at the university. We began the process of predictive analytics by creating a simple model comparing this year's admitted students to last year's admitted students. The goal is to predict the number of NEF beginning school at the start of the next fall term. The first and simplest model was based on ratios:

Predicted NEF enrollment for this year = (Actual enrollment for last year) * (Admitted student ratio)

In symbols:

$$Y_i = Y_{i-1} * \left(\frac{X_i}{X_{i-1}}\right)$$

where the Admitted Student Ratio $\left(\frac{X_i}{X_{i-1}}\right)$ is the ratio of the number of admitted students on a given date this year, divided by the number admitted students last year on the comparable date.

This model is crude for two reasons. It doesn't take into account admitted student counts for any years other than the preceding year. It also doesn't take into account the year-to-year changes in Yield, which is the proportion of admitted students who actually enrolled in the coming fall semester. Over time the yield rate at the university has been declining, from about 45% in the Fall of 2006 to about 36% in the Fall of 2012. Therefore, it takes additional admitted students, in comparison to last year, in order to maintain constant NEF enrollment.

Multiple regression estimation of NEF

An improved version of this model involves multiple regression. To predict the overall number of NEF in an upcoming fall semester, we can use multiple regression on a weekly basis during the academic year prior to the beginning of the students' college career. Our models take the form (for *i* = 39, 38, 37, ...):

$$Y_i = b0_i + b1_i * X1_i * b2_i * x2_i$$

The subscript i refers to the week counting backward from the start of the upcoming fall semester. Thus Y_{22} refers to the predicted number of NEF at a time point 22 weeks before the beginning of fall classes. Denoted by $X1_i$, we use the admitted students as of the end of the given week as a predictor variable for the fall enrollment of NEF. Thus $X1_{22}$ refers to the number of admitted students at a time point 22 weeks before the beginning of the semester. $X2_i$ refers to a year variable, where the current year is numbered 0 and previous years are numbered -1, -2, -3, etc. Our data for this model goes back to the Fall 2006 cohort, that is students applying during the academic year from Fall 2005 to Spring 2006. Thus we have seven years of observed data with the number of admitted students for the given week and the number of NEF enrolling the following fall semester. Because there has been a consistent decrease in yield over these years, a linear time trend variable $X2$ is included in the model.

For ease of computation and interpretation, the data is coded so that the current values of $X1$ and $X2$ are both 0. To make $X1 = 0$, we subtract the current number of admits from each observation, including the observations from previous years. Subtracting a constant from a variable doesn't change the estimated slope parameter for that variable; it only changes the intercept. By the same trick, the current value of $X2$ is coded to be 0, while the previous year is coded -1, the year before that -2, etc. With zeroes as the current values of both $X1_i$ and $X2_i$, the predicted value of Y_i is simply the estimated intercept b0i.students to last year's admitted students. The goal is to predict the number of NEF beginning school at the start of the next fall term.

Multiple Regression Model with Interaction, for Predicting the Number of NEF Enrolling the Next Fall

We complicate the previous model by adding an interaction term. Interaction in regression is useful when the effects of one variable depend on the values of another variable in the model. In our case, we create an interaction term by multiplying the coded number of admitted students, $X1_i$, by the coded year variable $X2_i$. Employing our earlier notation (for *i* = 39, 38, 37, ...):

$$Y_i = b0_i + b1_i * X1_i + b2_i * X2_i + b3 * X1_i * X2_i$$

With the interaction term we are able to capture deviations from the standard linear predictions based solely on $X1$ and $X2$. At present we run each of these models each week to update the forecast of the size of the coming Fall NEF cohort.

Logistic Regression (Logit) Model for NEF Enrollment

The weakness of the models in the last section is that they are at the cohort level of granularity, so that no individual student information is used. Only student counts are used in the forecasts, and only one observation is added to the data set with each successive year. Being able to use additional information on each student's (e.g. the student's high school GPA, ACT scores, FAFSA information, etc.) has the potential to enhance the predictive power of the forecast. We have found that students with mid-range GPAs and ACT scores are more likely to enroll at the university than are students with high GPAs and ACT scores. Furthermore, students sending in a FAFSA report for financial aid purposes are more likely to enroll than students not reporting any FAFSA information.

Another important predictor of enrollment is the time at which the university receives the application. For students admitted 250 or more days in advance of the fall semester, the yield is less than 30%, while for students admitted in the last 100 days in advance of fall semester; the yield is higher than 50%. A logistic regression model has been developed to account for these patterns in enrollment. This model uses the following equations:

Let i represent an index of the admitted NEF students for a given fall semester, $i = 1, 2, \ldots, n$ (the total number of admitted students).

Define $Y_i = 1$ if student i enrolls at the university, 0 otherwise.

$$Z_i = B0 + B1 * X1_i + B2 * X2_i + B3 * X3_i + \cdots$$

where subscript i represents the individual student index running from 1 to n.

$X1_i$, $X2_i$, $X3_i$, … represent observations on predictor variables such as high school GPA, composite ACT score, and Expected Family Contribution (EFC) determined by the FAFSA report submitted by the student.

$$P(Y_i = 1) = \frac{e^{Z_i}}{1 + e^{Z_i}}$$

The model has similar elements to a multiple regression model in estimating Z_i. The probability formula uses the inverse logit function (or the inverse of the canonical link function in generalized linear model terminology). The parameters $B0, B1, B2, \ldots$ are estimated from data collected over previous years. This data would include values for Y and all potentially related independent variables $X1, X2, \ldots$. The model has the advantage of supplying probability estimates at the individual student level, for whether or not the student will enroll in the coming fall semester. If these probabilities are summed for all admitted students in a given cohort, the result will estimate the number of these students that will enroll. This means that the logit model can be used in conjunction with the regression models described above, to confirm or call into question the forecasts of each model. When run during the 2012 year, we found that the regression models had a consistent positive bias, that is they had a tendency to produce forecasts larger than the eventual level of NEF enrollment. On the other hand, the logit models produced forecasts very consistent with the eventual level of enrollment. So it appears that the extra level of granularity in the data (student level versus cohort level) is beneficial for forecasting enrollment.

Because the level of granularity is at the student level, rather than at the cohort level, the logistic regression model has many additional uses outside of predicting the total number of NEF expected in the fall semester. For instance, the students with estimated probabilities closest to .50 are essentially "on the fence," and might be appropriate targets for extra contacts from the Admissions Office.

The estimated probabilities are easily updated from week to week, as additional students are added, and additional information becomes available. For example, FAFSA information about EFC is not available for most students in the early weeks of the recruiting year. Hence, because of missing data on that variable, the missing data dummy variables are employed for EFC, creating two variables, called for example EFC1 and EFC2, that indicate, respectively, whether or not the EFC value is known for a particular student and the value of EFC for that student when it is known. The values of EFC1 and EFC2 may change for various students on a week-to-week basis, which in turn will changes the estimates from the model each week. The same is true for other data which may be missing for a substantial number of student applicants, for example, high school GPA, high school percentage rank, and ACT Composite score.

When predicting student enrollment, the model is estimated from historical data, that is, data from previous years. Much student data (e.g. ACT scores, high school transcripts, student application information, housing applications, FAFSA information, recruiting event registration or participation) comes in over the course of the weeks leading up to the beginning of the fall term. For these data that arrive at unpredictable times, dates must be stored in the data warehouse for use in the modeling process. For example 30 weeks before the start of the term, all information that was not available 30 weeks before the start of that student's term must be coded as missing. This means that arrival dates must be known for all student data such as that listed above. The historical rec-

ord of a student's application process must include this type of chronological information.

Models

The following variables have been found to be strong predictors of student enrollment in a logistic regression model:

- Application Days before Term (the number of days before the start of fall term that the student was admitted)
- Housing Application Submitted (yes or no)
- Type of Application Submitted (paper or online)
- Student of Color (yes or no)
- Two or more racial group affiliations (yes or no)
- Distance (the number of miles from a student's high school to the university)
- Age category (19 and under, 20 and older)
- High School Percentage Rank (two dummy variables used)
- High School Region within the university's state
- EFC submitted (yes or no)
- Interest level indicated from ACT or contact postcard (two dummy variables used)
- Attendance at Circuit Fair recruiting event (yes or no)
- Attendance at Education state recruiting event (yes or no)
- Attendance at the university campus recruiting event (yes or no)
- Identification at High School visit by the university recruiter (yes or no)

Altogether, 18 variables would be used as predictors to represent this information. Most of these variables would eventually be available for every student being admitted. Several would have missing values for some students, especially for the early weeks in the recruiting year, necessitating double dummy variables for each. Essentially, a different model is estimated for every week of the year. Using the information that was available by week N of each of the years of data in the data cube, a model for that week is fit, with these variables. Typically the model for a given week is evaluated using either a forward selection or backward elimination technique for choosing the variables. There are many other candidate variables available from varied data collection sources, potentially allowing models with as many as 50 or more variables entering the model. However, not all of these variables would have coefficients that are statistically significant (i.e. significantly different from 0, based on the p-value for the associated t-statistic). Furthermore, it is often the case with complex models that the predictability of the model goes down as the number of variables included in the model increases. Though counter-intuitive, this feature prohibits us from using all variables at our disposal in a given model. Indeed, through some validation of models done in the past year, acceptable predictive power is achieved with fewer than the above 18 variables, although more validation is needed to confirm this result.

Error in a model of this type occurs naturally, in that all predictions for individual students are between 0 and 1, representing students with low probability of enrollment to high probability of enrollment. No fitted model is expected to perfectly predict the enrollment patterns of the students. Natural variation is expected in such a model, though the better fitting models will have predicted probabilities closer to the actual results 0 and 1. The data used to fit the model is called the "training data." The model is expected to have the least amount of error when applied to the very data on which it was fit. When the model is applied to new data, for example the next year's applicants for enrollment, we do not expect the fit to be as good. Optimizing a model to a given set of data naturally implies that it is less than optimal when applied to new data. Often the problem is that the model has been "overfit" to the data; it has used predictors which work especially well on the training data, but will not work as well on validation data. Typically a model can be cut down by examining which terms in the original model are least effective when applied to validation data.

Aside from expecting a generally weaker fit to the new data, a tool to assess the extent to which the new data will or will not conform to the old model is through cross-validation. This technique validates a model by splitting the original data into two parts: training data and validation data. Which observations go into which portion is done through a randomization process, to avoid bias in the selection of the training data. A common split percentage is to put 70% in training data, and the remaining 30% in validation. However, other splits (50-50, 60-40, 80-20) are also commonly used. Generally the training data set should be at least as large as the validation set. Larger amounts of data, that is larger numbers of student records, will allow for a more generous allocation into the training data. However, if data is scarce, then few observations can be spared for validation, as the majority will be needed for training the model.

If several competing models are selected on the basis of fitting them to the training data, they may each be tested on the validation data, in order to choose the best. It is not necessarily the case that the best fitting model for the training data will turn out to be the best fitting model

on the validation data, especially if one or more other models are nearly as good on the training data. Use of cross-validation allows us a preview of how well each model may do on new data to be obtained in the next year, assuming conditions for the collection of next year's data are similar to this year and previous years on which the model was built.

When a logistic regression model has been selected for a given week, based on training and validation data from previous years, it can then be applied to the current year's data to predict student enrollment. Each student will then have a predicted probability of enrollment, and these probabilities can be summed for all student applicants to estimate the number of students that will enroll from those that have been admitted thus far. Of course, there will be other students that apply for admission after this particular week, and so they would be included in predictions for some future weeks, though they are not included for this particular week. Thus, the predicted number of new students that will enroll increases from week to week with a logit model, as the prediction includes more and more admitted students each week. To have an idea of the total new student enrollment in the fall, one also needs to estimate the enrollment from late applications. This estimate may come from previous year's data, for instance, by finding the average number of students enrolled per year who applied after this given date in the recruitment cycle. Or regression models may be attempted, if the number of late applicants is related to the number of earlier applicants in some way.

Decision Trees

Another method useful in predictive analytics is the decision tree. The set of all cases are split according to some value in one of the predictor variables. The predictor and the splitting value are chosen to optimally separate the values of the dependent variable, whether or not students enroll, for example. In the tree model most recently used to predict university student enrollment, the first split in the data was according to the distance the student's high school was from the university. This split was the following:

- Group (1): If Distance from the university < 48.1 miles, then predict student will enroll;
- Group (0): If Distance from the university >= 48.1 miles, then predict student will not enroll.
- This split is followed by other splits to optimally separate cases. Variables can be used multiple times for splits in various branches. The next split in the student enrollment tree is in Group (1):
  - Group (11): For Group 1, if Application Days Before Term < 167 days, then predict student will enroll;
  - Group (10): For Group 1, if Application Days Before Term >= 167 days, then predict student will not enroll.
  - Another split can be made in Group (0), which just happens to use the same variable:
  - Group (01): For Group 0, if Application Days Before Term < 139 days, then predict student will enroll;
  - Group (00): For Group 0, if Application Days Before Term >= 139 days, then predict student will not enroll.

More branches can be formed until the gain from such splitting no longer adds significant predictive power. At this point the tree is complete.

As with the models mentioned earlier, it is best to break the original data into training and validation sets. So a tree might be fit on 70% of the original data, and then applied to the 30% validation set for confirmation. The validation fit will generally not be as good, and some pruning of the original tree might be in order, to avoid the problem of over-fitting. This cross-validation is especially effective for tree models, as over-fitting is to be expected in the original tree. Some modeling software offers the user 5-fold or 10-fold cross-validation options, whereby consecutive random portions of the data are held out and the model is refit to all data not being held out. If the random portions are 20%, then five such holdout samples are created, and the model is fitted five times to the corresponding 80% portions that remain. This would be called 5-fold cross-validation. An overall measure of goodness of fit is applied to determine which variables and splits should remain in the model. Typically the overall model will be smaller than each of the five initial models.

DISCUSSION AND CONCLUSIONS

When undertaking a very complex project such as implementing BI in a higher education institution obtaining complete success is a most elusive goal. However, in the authors' case many successful steps occurred during the development and implementation of the BI strategy as well as several failures. It is hoped that the failures can be re-evaluated and serve as lessons learned to guide future adjustments to the system. The remainder of this section shall delineate what was done right and done wrong during the development and implementation within the authors' system.

What was done right (Proficiency)

Perhaps one of the most important steps was to obtain executive level support and leadership throughout the analytics design process. This was crucial in being able to get buy in from the various departments across the campus. This high level support led to the development of strong and flexible application of analytics within the system. More specifically, the authors were able to:

- Leverage analytics to redirect academically underprepared university applicants to alternative remedial options at a partnering institution
- Leverage analytics to provide scholarship incentives for high academically performing applicants with an EFC (Expected Family Contribution) of $0
- Select appropriate skill sets for staffing the analytics initiative
- Remove low impacting variables during the regression modeling process to improve the prediction validity of the remaining variables
- Leverage statistical models that fit well with the authors' University's goals
- Correct business process according to the analytic model results
- Normalize by rescaling variables to a standard scale (i.e. high school GPA)
- Address the problem of missing data
- Leverage the use of dummy variables (i.e. ACT scores)
- Impute values
- Leverage multiple regression with interaction for projecting enrollment
- Create a logit (logistic regression) model for student level analysis
- Validate the logit model against the multiple regression model
- Create a decision tree model and then employed cross validation techniques.

Similarly, much success was also obtained in the area of data architecture and validation. This was most encouraging as sound data is the foundation for any BI system. More specifically, the authors were able to:

- Select the appropriate Warehouse design
- Leverage the automation potential inherent in the modular design aspects of the Inman approach
- Reduce the self-service learning curve by presenting a relational enterprise data source
- Create an architecture that retained historical data at the design phase of the data warehouse initiative therefore providing more longitudinal options to the analytic modelers
- Transfer ambiguous data into the bit-bucket
- Require a relational constraint validation prior to loading source data into the data warehouse
- Provision executive level administration to drive institutionally adapted refresh rates for all Data Mart tables.

What was done wrong (lessons Learned)

There were also a number of challenges in developing the system. Once again these problems are being addressed and it is expected that other problems will surface as the system matures. One must accept that BI development is a longitudinal problem and feedback loops as well as remedial procedures are necessary to keep the system viable over time.

Because of the complexities of integrating data from various sources problems occurred in the development of the data warehouse. Some of the problems observed follow:

- Creation of one-off static data sets rather than dynamic operationalized warehouse sources for statistical modeling
- Implement a Data Warehouse prior to establishing enterprise data governance which lead to:
- Not correcting business practices to ensure relationally accurate operational data
- Poor data quality resolution processes restricted vendor selection options
- Inconsistent corrective action by data owners in response to errant data claims
- Lack of comprehensive integration with other department data initiatives.

Perhaps the most success limiting set of problems, were related to lack of understanding on the part of the end-users in regard to what it takes to implement a successful BI system. Solid design and effective data architecture are the cornerstones of a BI system. For end-users these are often developed behind the scenes and become transparent. Therefore, the end-user often focuses primarily on the reports generated as a function of the BI system. End-user concerns are often related to reliability timeliness, ease of use and correctness of those reports. Once again if the data is not properly developed one will be stuck with the garbage-in/garbage-out model. There is a fine line in giving the end-user some preliminary reports early on to validate the process while not generating final reports that will be inaccurate because the data sources

are not yet mature. The problems the authors observed in this area follow:

- Executive team placed a disparate degree of resources in report generation, while short changing data warehousing and inadequately supporting data cleansing processes
- Overly specific and narrowly focused self-service BI initiatives resulting in reporting solutions failing to meet changing end user expectations.

In summary, BI can make a huge difference in regard to the efficiency in which an institution of higher education operates. However, it is a resource intensive process and requires buy in at all levels, particularly the executive level. Assembling a top notch team is paramount if an efficient system is to be developed. Too often the players don't understand the complexities and rush to generate the reports before a sound foundation is in place. If possible, the Inmon approach [21] should be considered and implemented if at all practical. The more planning that takes place before implementation, perhaps the greater probability of success and may even end up being quicker in the long run!

REFERENCES

[1] Ambler, S., "Whence data management," *Dr. Dobbs Journal*, October 5, 2006.

[2] Andriole, S., "The collaborate/integrate business technology strategy," *Communications of the ACM*, 49(5), pp. 85-90.

[3] Ang, J. and Teo, T. S. H., "Management issues in data warehousing: insights from the Housing and Development Board," *Decision Support Systems*, 29 (1), 2000, pp. 11-20.

[4] Apte, C., Liu, B., Pednault, E and Smyth, P., "Business Applications for Data Mining," *Communications of the ACM*, 45(8), 2002, pp. 49-53.

[5] Argotte, l., Mejia-Lavalle, M and Sosa, R., "Business Intelligence and Energy Markets: A Survey," *Proceedings of the 15th International Conference on Intelligent System Applications to Power Systems (ISAP)*, November 8-12, 2009.

[6] Ariyachandra, T. and Watson, H., "Key organizational factors in data warehouse architecture solutions," *Decision Sup-port Systems*, 49 (2), pp. 200-212.

[7] Balkan,S. and Goul, M., "A Portfolio Theoretic Approach to Administering Advanced Analytics: The Case of Multi-Stage Campaign Management," *Proceedings of the 44th Hawaii International Conference on System Sciences (HICSS)*, January, 2011.

[8] Bao-sheng, L. and Xin-quan, G., "Study on Predictive Model of Customer Churn of a Mobile Telecommunications Company," *Proceedings of the 4th International Conference on Business Intelligence and Financial Engineering*, October, 17-18, 2011.

[9] Berg, B., "Predictive Modeling: A Tool, Not the Answer", *University Business*, July/August, 2012.

[10] Chaudhuri, S., Dayal, U. and Narasayya, V., "An Overview of Business Intelligence Technology," *Communications of the ACM*, 54(8), 2011, pp. 88-98.

[11] Chiang, R., Goes, P. and Stohr, E., "Business intelligence and analytics education, and program development: a unique opportunity for the information systems discipline," *ACM Transactions on Management Information Systems*, 3(3), 2012.

[12] Cupoli, P., Devlin, B., Ng, R. and Petschulat, S., "ACM Tech Pack on Business Intelligence/Data Management," *ACM*, 2012, pp. 15.

[13] Denny, P., "BiLog: Best practices…Matching maximo BI tools and maximo users", https://www.ibm.com/developerworks/community/blogs/a9ba1efe-b731-4317-9724-a181d6155e3a/entry/bilog_best_practice_matching_maximo_bi_tools_and_maximo_users13?lang=en, March 7, 2013

[14] Duggin, J., Cetintemel, U., Papaemmanouil, O. and Upfal, E., "Performance prediction for concurrent database workloads", *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, pp. 337-348.

[15] Execution MiH., "Refresh Frequency of a Data Warehouse: What should be the refresh rate of a data warehouse?" http://www.executionmih.com/q/refresh-frequency-of-a-data-warehouse.php, 2012.

[16] Ferranti, J., Lnagman, M., Tanaka, D., McCall, J. and Ahmad, A., "Bridging the gap: Leveraging business intelligence tools in support of patient safety and financial effectiveness," *Journal of the American Medical Informatics Association*, 17(2), 2011, pp. 136-143.

[17] Florescuand, D., "An extensible framework for data cleaning," *Proceedings of the 16th International Conference on Data Engineering*, 2000, pp. 312.

[18] Goldstein, P., "Academic Analytics: the Uses of Management Information and Technology in Higher Education," *ECAR Key Findings, EDUCAUSE*, December, 2005.

[19] Graf, S., Ives, C., Lockyer, L., Hobson, P. and Clow, D., "Building a data governance model for learning analytics," *Proceedings of the 2nd Interna-*

tional Conference on Learning Analytics and Knowledge, ACM, New York, 2012, pp. 21-22.

[20] Guster, D. C., & Brown, C., "The Application of Business Intelligence to Higher Education: Technical and Managerial Perspectives," *Journal of Information Technology Management*, 23(2), 2012, 42-62.

[21] Inmon, W., *Building the Data Warehouse*, John Wiley & Sons, New York, 1992.

[22] Isik, O., Jones, M. and Sidorova, A., "Business Intelligence (BI) Success and the role of BI capabilities," *International Journal of Intelligent Systems in Accounting and Finance Management*, 18(4), 2011, pp. 161-176.

[23] Kimball, Ralph; et al., *The Data Warehouse Lifecycle Toolkit*, New York: Wiley, 1998.

[24] Krupa, M., "HR business intelligence Apps-don't do the kitchen sink method", *InfoBox: a Tidy Package of HR and Technology Information, Insight , Wit and Wisdom*, http://infoboxinc.com/hr-business-intelligence-apps-dont-do-the-kitchen-sink-method/, July 11, 2010.

[25] Laun, J., "Data Mining and Applications in Higher Education," *New Directions for Institutional Research*, 113, Spring 2002.

[26] Lupu,A., Bologa, R., Lungu, I. and Bara, A., "The Impact of Organizational Changes on Business Intelligence Pro-jects," *Proceedings of the 7th WSEAS International Conference on Simulation, Modeling and Optimization*, 2007, pp. 415-419.

[27] Marjanovic, O., "The next stage of operational business intelligence: Creating new challenges for business process management," *Proceedings of the 40th Annual Hawaii International Conference on System Science, IEEE Computer Society*, 2007, pp. 215c.

[28] Mehta, A., Gupta, C. and Dayal, U., "BI batch manager: a system for managing batch workloads on enterprise data-warehouses," *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, pp. 640-651.

[29] Moss, L. and Atre, S., *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications*, Addison-Wesley Longman, Boston, 2003.

[30] Natsu, J., "Advanced Analytics: Helping Educators Approach the Ideal", *eSN Special Report, eSchool News*, 2010, pp. 17-23.

[31] Neubock, T., Neumayr, B., Rossgatter, T., Anderlik, S. and Schrefl, M., "Multi-dimensional navigation modeling using BI analysis graphs," *Proceedings of the International Conference on Advances in Conceptual Modeling*, Springer Verlag, Berlin, 2012, pp. 162-171.

[32] Pucher, M., "Management is not a science: It is an art. Welcome to Real World IT", http://isismjpucher.wordpress.com/category/business-intelligence-2/, 2010.

[33] Ramakrishnan, T., Jones, M. and Sidorova, A., "Factors influencing business intelligence (BI) data collection strategies: An empirical investigation", *Decision Support Systems*, 52(2), 2012, pp. 486-496.

[34] Ramamurthy, K. R., Sen, A. and Sinha, A. P., "An empirical investigation of the key determinants of data warehouse adoption," *Decision Support Systems*, 44 (4), 2008, pp. 817-841.

[35] Ranjan, R., "Business intelligence: Concepts, components, techniques and benefits," *Journal of theoretical and Applied Information Technology*, 19(1), 2009, pp. 60-70.

[36] Sahay, A. and Mehta, K., "Assisting Higher Education in Assessing, Predicting and Managing Issues Related to Stu-dent Success: A Web-based Software Using Data Mining and Quality Function Deployment," *Academic and Business Research Institute Conference Proceedings*, Las Vegas, 2010.

[37] Scannapieco, M., Mirabella, V., Mecella, M. and Batini, C., "Data quality in e-business applications," *Lecture Notes in Computer Science*, 2512, pp. 121-138.

[38] Schonberg, E., Cofino, T., Hoch, R., Podlaseck and Spraragen, S., "Measuring Success," *Communications of the ACM*, 43(8), 2000, pp. 53-57.

[39] Segev, A., Shao, S. and Zhao, J., "A data analysis model for business intelligence", *International Journal of Internet and Enterprise Management*, 1(1), 2003, pp. 7-30.

[40] Shmueli, G. and Koppius, O., "Predictive analytics in information system research," *MIS Quarterly*, 35(3), 2011, pp. 553-572.

[41] Themistocleous, M. and Irani, Z., "Evaluating and Adopting Application Integration: The Case of a Multinational Petroleum Company," *Proceedings of the 35th Hawaii International Conference on System Science*, 2002.

[42] Ulanov, A., Simanovsky, A. and Krohn, M., "Personalized report creation for business intelligence", http://persdb2012.cs.umn.edu/papers/8.Ulanov-PersDB12.pdf, 2012

[43] Van Barneveld, A., Arnold, K. and Campbell, J., "Analytics in Higher Education: Establishing a Common Language," *EDUCAUSE Learning Initiative Paper*, Number 1, 2012.

[44] Williams, S. and Williams, N., "The Business Value of Business Intelligence," Business Intelligence Journal, 8(4), 2003.

AUTHOR BIOGRAPHIES

Christopher G. Brown is the Lead Business Intelligence Data Architect for Saint Cloud State University. His work focuses on the modeling relational and multidimensional databases for higher education administrative decision making. Christopher's data architecture emphasizes n-tiered modular automated data flow processing. Current projects include modeling enrollment and retention predictive analytics.

Dr. Dennis Guster is a Professor of Computer Information Systems and Director of the Business Computing Research Laboratory at St. Cloud State University, MN, USA. His interests include network design, network performance analysis and computer network security. Dennis has 25+ years of teaching experience in higher education and has served as a consultant and provided industry training to organizations such as Compaq, NASA, DISA, USAF, Motorola, and ATT. He has published numerous works in computer networking/security and has undertaken various sponsored research projects.

Erich P. Rice is a graduate student in the Master of Science in Information Assurance (MSIA) program at St. Cloud State University (SCSU). He has been the graduate assistant for the Center for Information Assurance Studies at SCSU for the past year. He has a Law Degree from William Mitchell College of Law, and is interested in legal issues pertaining to Information Assurance.

Dr. David Robinson is a Professor of Statistics, in the Department of Mathematics and Statistics at St. Cloud State University, MN, USA. He has more than 34 years of teaching experience at the university level, and has also served as statistical consultant on numerous projects for universities, business, and industry. His current interests involve application of statistical tools to data mining and analytics. Other interests are public opinion research, Monte Carlo simulation techniques, and nonparametric statistics. He has worked collaboratively to author numerous conference and journal papers, and one book.