AN ANALYTICS SOLUTION FOR RETAIL EMPLOYEE TURNOVER WORKFORCE MANAGEMENT

Akhilesh Karumanchi[1], Mayank Gupta, Matthew A. Lanham
Purdue University Krannert School of Management
403 W. State Street, West Lafayette, IN 47907
akaruman@purdue.edu, gupta363@purdue.edu, lanhamm@purdue.edu

ABSTRACT

This study integrates the three domain of business analytics: descriptive, predictive, and prescriptive to develop a workforce management turnover solution that could be used by retailers. First, we investigate and summarize relationships among features to identify potential cause-and-effect relationships. Second, we build and evaluate predictive models to estimate the probability that a team member will leave a retailer within a future planning horizon. Lastly, a theoretical decision model is formulated that provides guidance in how a practitioner might use the prediction outputs for future decision or policy making (e.g. raise decisions, firing decision, etc.). This study is novel in the framework we propose in how to integrate analytics to support the retail employee turnover problem. Most research we have studied discuss employee turnover on theoretical grounds, rather than providing analytical decision-support solutions which are vast in other business verticals. Using data from a local retailer we develop a working framework that provides guidance to human resource professionals in how descriptive, predictive, and prescriptive analytics can be aligned to address employee turnover.

[1] Corresponding author (Phone: 231-203-4060)

INTRODUCTION

Retailers can achieve success when they retain and reward their best people. Employee turnover is costly if the employee who is leaving the company is a high performer. High performers in the company are those people whom have been recognized or evaluated by their superior's or peers as contributors towards the success of the company. Nowadays, the big challenge for HR Managers is to retain the best people by developing policies that keep them market competitive, and employees incentivized not to leave, while having the ability to meet or exceed the expectations of their customers.

The purpose of the retailer is to provide consumers a convenient avenue to acquire the products they desire from many manufacturing and service providers. Studies have shown that countries having the greatest economic and social progress have been those having a strong retail footprint (Miller 2009). In the United States, approximately 15% of all jobs are in retailing (Dunne, Lusch et al. 2013). Walmart, the world's largest retailer employs more than two million people worldwide (Dunne, Lusch et al. 2013).

Most industry work on retail analytics comes from a supply chain perspective, which focuses on the various stages of getting and presenting products directly to their customers. The key here is knowing your customer. For example, retailers invest much time identifying consumer buying patterns which can provide assortment and pricing insights, but also provide guidance in which coupons to offer customers with the goal of gaining more of their business. Marketing departments will regularly analyze transaction log data, in-store checkout wait times, and store foot traffic to develop modified marketing strategies to better serve their customers (Brust 2013). Some consumers are willing to provide personal details about themselves if it provides them some benefit. Some claim that employing personalized marketing to such individuals can boost sales by ten percent and provide five to eight times the return on investment (Hoffman and Fodor 2010).

Retailers use point-of-sale (PoS) systems frequently to capture and store precise information about what was purchased, when it was purchased and whom purchased it. Many retailers will also have customer loyalty programs to help increase the transparency of these purchases by having a unique customer profile id associated with each purchase. PoS data contain transaction data such as time and place of transaction, products purchased, if coupons were used, and type of payment, such as cash or credit card. Some interesting case studies investigating PoS data can be found by (Cadez and Smyth 2001, Mladenic, Eddy et al. 2001, Shashanka and Giering 2009).

This remarkable amount of data collected and analyzed are also often used for category planning decisions such as shelf layout and supplier selection decisions. Category planning entail a series of hierarchical decisions such as category sales planning, assortment planning, shelf space planning, and in-store logistics planning encompass master category planning (Hübner and Kuhn 2012). We found there is a tremendous amount of analysis here to decide how to lay out a store and fill it with the best possible set of products. These problems also lead to the literature rich area of pricing.

If the retailer is doing all these correctly, the next thing a retailer might want to know is which stores are performing better than others. For example, clustering stores with similar profiles may

be used by management to identify "underperforming" stores in their respective cluster to identify possible underperformance causes and potentially provide those stores additional resources or employee training.

The interesting thing in all this data collection and analysis is this is where the advanced analytics seem to stop. We have found that a retailer will invest much to understand their customers, but little to understand their employees. Per Deloitte' s Global Huma Capital Trends 2014 report, just 14% of the Human Resource departments use data analytics to perform their jobs (Feffer 2014). Employee centric analytics is negligible compared to operations (77%), sales (58%), and marketing (56%) as shown in Figure 1.

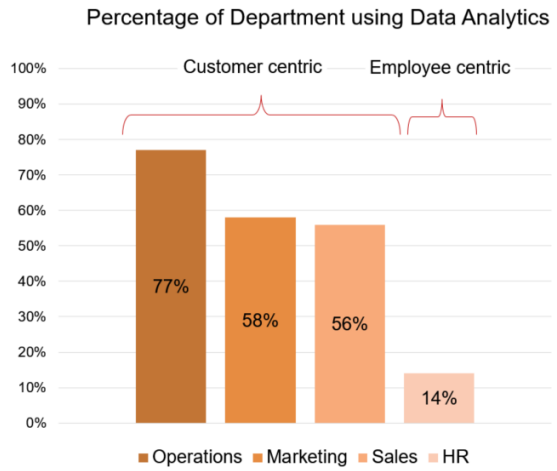

Figure 1: Percentage of departments using data analytics

We posit that the reason for a lack of data analytics among professionals in Human Resource (HR) departments has to do with a deficiency of training in analytical skills compared to other functional business units. We believe HR areas could collaborate with other units within their organization to help better them better utilize the information about their employees as depicted in Figure 2. This motivation led us to develop an analytical framework that HR areas could build upon to better understand and support their decision and policy making within the firm. This solution could be modified and used by HR decision makers using their own employee data, having their own drivers of turnover, and own specific business constraints.

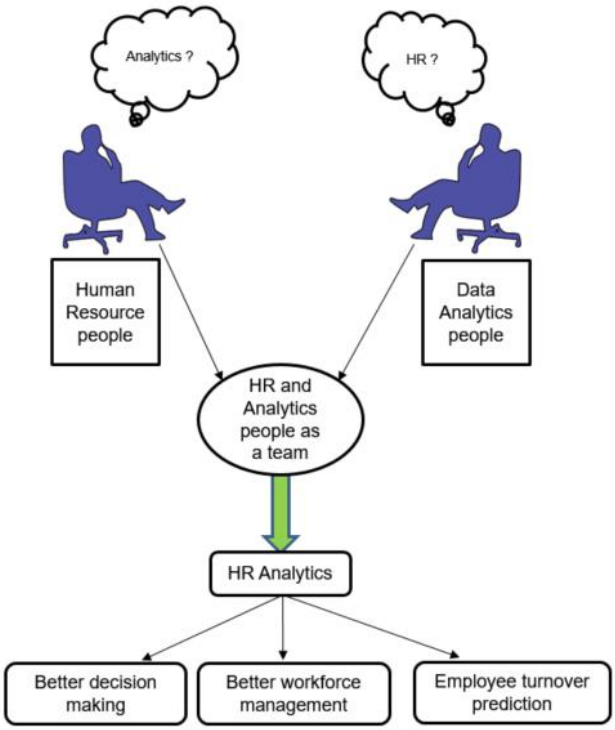

Figure 2: Bringing HR together with other functional units that regularly use data analytics

We organize this paper by first reviewing and organizing the academic literature on various topics relating to employee turnover. Second, we discuss the data used in our study to help guide the development of a prototype framework. Third, we outline the methodology employed to gather insight from data. Fourth, we explain the predictive models we explored to estimate the probability that a team member will leave over some future planning window. Fifth, we discuss the results of our predictive models. Lastly, we develop a decision model that incorporates information gained from our predictive models with the objective of helping a human resource team manage their retail workforce.

LITERATURE REVIEW

We critically reviewed prior research to provide a solid foundation and clear perspective to guide our research in the area of employee turnover. Strategic HR researchers are investing much effort to understand if it is the HR policies themselves or the performance of the company which affects employee turnover (Collins and Clark 2003, Hatch and Dyer 2004).

Employee turnover studies in the 1970s and early 1980s have the basic tenet that job satisfaction was the main reason for the employee turnover, but later on it became other job alternatives. Graen, Liden et al. (1982) found that the quality of the leader–member exchange relationship predicted employee turnover, and (Pfeffer 1985) argued for the importance of demographic fit. During this time, researchers also attempted to identify the consequences of turnover, these early studies primarily focused on increased costs and organizational performance decrements following turnover (Mirvis and Lawler 1977, Price 1977, Dalton and Todor 1979, Staw 1980, Mobley 1982).

In the mid-1980s into the early 1990s, most of the employee turnover research was based on exhaustion and stress factors that are related to organizational culture, employee relationships in the environment, organizational reward system, group cohesion, gender composition and demography. In one study, pay dispersion defined as "*the amount of pay inequality within an organization's pay system*", predicted turnover among university administrators such that turnover was lower at institutions with more compressed pay structures (Pfeffer and Davis-Blake 1992). O'Reilly, Chatman et al. (1991) found that variables such as coworkers and boss were also factors, and that employees whose personal values did not align with the organization's values (e.g. low person–organization fit) had a greater likelihood to turn over after 20 months of tenure. Further, the establishment of mentoring relationships was shown to reduce employee turnover intentions (Viator and Scandura 1991, Payne and Huffman 2005). McPherson, Popielarz et al. (1992) investigated employee turnover from a social network perspective and identified that employees with more connections within an organization's social network were less likely to leave. For example, emotional exhaustion and job insecurity were found to be positively related to turnover intentions (Jackson, Schwab et al. 1986, Ashford, Lee et al. 1989). Lee and Mitchell (1994) developed a new theory and model regarding the turnover process. Drawing from image theory, they proposed that turnover decisions are not always the result of accumulated job dissatisfaction and may sometimes occur without much deliberation at all. In other words, sometimes things this just happen such as an unexpected new job opportunity.

Different types of organizations and industries face very different average turnover rates. Even across firms in the same industry, turnover rates can vary widely. Thus, it is important for firms to differentiate between avoidable (i.e. understanding the causes) and unavoidable turnover (Barrick and Zimmerman 2005). In retail for example, turnover rates can average around 30% over any planning horizon. This will affect the individual store level performance in terms of both sales and profit (Kacmar, Andrews et al. 2006).

Employee turnover has more implication on the employee than the company which he/she left, because the employee must get accustomed to new situations, new working styles, and new people. Some items these studies found were that employee turnover depends on various factors like demographics and opportunities (Hom, Roberson et al. 2008).

In response, managers have implemented human resources policies and practices to actively reduce avoidable and undesirable turnover (Michaels, Handfield-Jones et al. 2001, Fulmer, Gerhart et al. 2003, Kacmar, Andrews et al. 2006, Hom, Roberson et al. 2008).

Today, employee turnover problem can be analyzed with the help of data analytics. Measuring and collecting the right employee data can provide better insights about what drives people to decide if they should leave an organization. As business analytics continues to be incorporated in other business domains such as Marketing, Operations, and Strategy policy making, we posit empirical-based models can also be used to better support HR policy making.

Studies	Motivation for the research	Result of the research
(Mirvis and Lawler 1977)	To Study factors for the employee turnover	Decrease in the organizational performance is a major factor for employee turnover
(Price 1977)	To study how can employee turnover be positive	To categorize the turnover as a positive or negative phenomenon seems somewhat short-sighted. It certainly

		has both positive and negative effects on the organization's performance
(Dalton and Todor 1979)	To study whether employee turnover can be positive	It depends on the level of turnover and the employees who are turning out. It can be positive if they lose unproductive employees
(Staw 1980)	To analyze the post-effects of the employee turnover	Consequences of the turnover will vary from company to company and it should be evaluated using descriptive inquiry
(Pfeffer 1985)	To study the effect of organization demography's implications on management	The organizational demography helps us to manage careers of the individuals and their needs based on their demographic characteristics
(Jackson, Schwab et al. 1986)	To understand the employee burnout phenomenon	Emotional exhaustion is the predominant employee burnout component
(Lee, Ashford et al. 1990)	To study the workers' satisfaction, performance and somatic complaints	The performance of the employees increases when they have a high degree of perceived control
(Viator and Scandura 1991)	To study the mentor-protégé relationship	It is difficult to measure the variation of quality done between mentored and non-mentored employees over the long time
(Mobley 1992)	To understand the causes, consequences of the turnover	Workplace atmosphere changes when employees frequently turnout and it affects the performance of the organization
(McPherson, Popielarz et al. 1992)	To study the dynamic behavior of the voluntary groups	More contacts a person has inside the group then more is the probability that he will stay as the long-term member.
(Pfeffer and Davis-Blake 1992)	To study the turnover among the college administration	An individual's position in the salary structure and level of dispersion in the structure jointly affects the turnover
(Handfield-Jones, Michaels et al. 2001)	To study role of leader in American retail organizations	Establishing talent standards and managing the talent of the employment is the crucial trait for a leader in retail organization in America
(Fulmer, Gerhart et al. 2003)	To study the relationship between the *Great place to work* and firm performance	The performance of a company varies as per the variation of employee attitudes and management change
(Hatch and Dyer 2004)	To study the human capital management as a competitive management	The cost advantages that can be attributed to human capital are sustainable because human capital is costly to imitate
(Payne and Huffman 2005)	To study the impact of organizational commitment on the employee turnover	The affective commitment and continuance commitment would mediate the relationship between mentoring and turnover behavior
(Barrick and Zimmerman 2005)	To study effective selection to avoid the voluntary turnover	It concludes that relevant bio-data and work-related dispositions assessed prior hiring any candidate can predict the voluntary turnover
(Kacmar, Andrews et al. 2006)	To study how costly is the turnover	Turnover will affect the individual store level performance in terms of both sales and profit
(Huselid and Becker 2006)	To study the human resources analytical literacy	If organizations can increase the analytical literacy of their HR professionals, then it can help them take strategic decisions on managing their workforce
(Holtom, Mitchell et al. 2008)	To study the turnover and retention of the employees	Turnover and retention can be positive and negative depending on the efficiency of the employee who left or who didn't leave the company
(Hom, Roberson et al. 2008)	To predict the employee turnover in corporate America	The employee turnover depends on various factors like demographics, opportunities, etc.
(Fidalgo and Gouveia 2012)	To measure the employee turnover impact on the organizational knowledge management	Conceptual map was created to demonstrate the prominence of the Knowledge management in the company and what are the organizational changes that are required

Table 1: Review of employee turnover studies

In Table 2 we identify studies where predictive modeling was performed to predict employee turnover. The studies showed that data mining techniques will help the human resource practitioners to speed up their process with much better efficiency. Several propose the idea of using data mining techniques in human resource practices by demonstrating its prominence in

leveraging decision making with their frameworks. However, most do not provide any empirical investigation. In our study, we use data mining techniques (i.e. descriptive and predictive analytics) on employee data, but then integrate those insights into an optimization model which considers many practical constraints of the real world.

Studies	Motivation behind research	Methodologies used/proposed to use	Results/ Proposed Frameworks
(Mishra, Lama et al. 2016)	Existing huge amount of data about employees and HR practices	Descriptive Analytics, Predictive Analytics, Prescriptive Analytics	Generalized decision making model without testing
(Fatima and Rahaman 2014)	A problem to manage faculty staffing in their university	C 4.5 Algorithm, Association rule, K-Nearest Neighbor, Apriori Algorithm.	Cyclic decision model but without testing
(Sadath 2013)	Connecting Human Resource Management to Knowledge Management	Association rule, Clustering, Prediction, Classification	Implement knowledge management programs for competitive advantage
(Mishra and Lama 2016)	To optimize performance and practice better return on investment for optimizations	-	-
(Feffer 2014)	HR typically lags in using data analytics	-	HR professionals should be comfortable in using data tools
(SEBT and YOUSEFI 2015)	Data mining being just limited to statistical analysis in HR	Statistical analysis using regression method, CART (ordered), CART (Towing), C 5.0	Data mining can give deep insights than simple statistical analysis

Table 2: Summary of studies using predictive modeling for employee turnover

Table 3 summarizes the studies suggesting which data predictive modeling methodologies could be used for better decision making. We implement some of these in our study as well as others that are popular in classification-type problems.

	Linear Regression Method	Logistic Regression	CART	Association rule	Apriori Algorithm	K-Nearest Neighbor	Random Forest	Support Vector Machines
(Mishra and Lama 2016)	-		-	-	-	-	-	-
(Fatima and Rahaman 2014)	-		✓	✓	✓	✓	-	-
(Sadath 2013)	-		-	✓	-	-	-	-
(Mishra, Lama et al. 2016)	-		-	-	-	-	-	-
(Feffer 2014)	-		-	-	-	-	-	-
(SEBT and YOUSEFI 2015)	✓		✓	-	-	-	-	-
Our Study	-	✓	✓	-	-	✓	✓	✓

Table 3: Comparison of methods used in the literature to predictive employee turnover

DATA

The data investigated in this study came from a regional retailer in the United States. There were certain store employees that began realizing a higher turnover than planned. We refer to these position titles as Job A, Job B, Job C, and Job D. These positions have similar hourly rates pay distributions and do not necessarily report hierarchically to each other. All jobs were present in each region of the 14 regions provided. In addition to these features, we obtained other variables

that measured employee performance, store performance, wage information, and market information as displayed in Table 4. The data set consisted of 1000 observations of employees whom have worked for the company for at least two years, where half the records indicated the employee left while the other employees were still employed.

Variable	Type	Description
Region	Categorical	Region id
Position Title	Categorical	Job title of that employee (Job A, Job B, Job C, Job D)
Market Group	Categorical	Market group the employee is associated to based on the store they work at
Pflag	Categorical	Hourly rate less than the market group range = -1, Hourly rate within the market class range = 0, Hourly rate more than the market class range = 1,
Pflag_below	Categorical	If hourly rate less than the market group range = 1, otherwise = 0
Pflag_above	Categorical	If hourly rate more than the market group range = 1, otherwise = 0
EmployeePerfPrevYr	Categorical	Employees annual performance review indicator for the previous year (Superior, Good Performer, Needs Improvement, Unsatisfactory)
EmployeePerfCurrentYr	Categorical	Employees annual performance review indicator for the current year (Superior, Good Performer, Needs Improvement, Unsatisfactory)
Hourly Rate	Numeric	Hourly wage rate ($)
Pnormal	Numeric	Normalized hourly rate according to the market class range given (0 indicates market average)
GroupPerfPrevYr	Numeric	Store annual performance for the previous year on a scale of 50-100
GroupPerfCurrentYr	Numeric	Store annual performance for the current year on a scale of 50-100
Left	Categorical	If employee has left = 1, If employee is still employed = 0

Table 4: Data Table

METHODOLOGY

The methodological framework we propose includes using descriptive, predictive, and prescriptive analytics together as shown in Figure 3. Retailers can obtain whatever data is available about their employees, job functions, etc. to identify and estimate cause-and-effect relationships of the drivers of employee turnover.

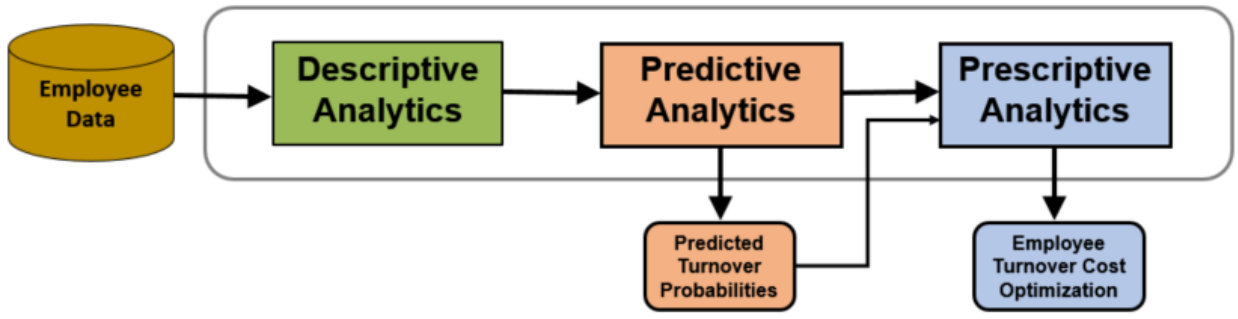

Figure 3: Methodology diagram

Descriptive analytics

Descriptive analytics refers to understanding what has happened in the past. One objective here is to try and identify cause and effect relationships of the business problem. For example, what if an employee is given a raise that is an average rate within their given market, how does this affect the likelihood of them leaving the company over the next planning horizon. Exploratory data analysis (EDA) which often entail data visualizations, statistical summaries, and correlation analysis can help in forming hypotheses about the causes of employee turnover.

Predictive analytics

Predictive analytics refers to understanding what will happen in the future. This domain helps identify and estimate the effect of each variable with regard to the response. In our study, our response is a binary variable left or not left the company. The independent variables are the remaining variables described in Table 4. We build and evaluate six different predictive models to identify these drivers and estimate their effects on turnover. The predictive models were built using a 70/30 train/test partition. The training set is used to build the model, while the test set allows us to gauge generalizability on future observations. Comparing the statistical performance measures (e.g. overall accuracy, AUC, etc.) allows one to identify if a model has overfit to the training data, and will perform poorly at identifying whom will leave or stay in the future.

Models were assessed using traditional binary classification statistical performance measures, such as area under the curve (AUC), accuracy, sensitivity, and specificity. All of these measures are generated from a confusion matrix depicted in Figure 4. First, accuracy calculates how well your model can classify those employees whom left the company versus those that did not leave the company. Those are the green colored cells in the table divided by the total observations (TP + TN)/Total. Specificity measures how well a model performs at identifying true leavers among the set that left the company, TN/(FP+TN). Similarly, sensitivity measures how well a model performs at identifying employees that actually left the company, TP/(TP+FN).

		Employee Actually Left Company		
		Yes	No	
Predicted Employee Left Company	Yes	*TP*	*FP*	TP + FP
	No	*FN*	*TN*	FN + TN
		TP + FN	FP + TN	Total

Figure 4: Confusion matrix

The accuracy, specificity, and sensitivity statistics are derived from one confusion matrix based upon a naïve predicted probability cutoff value of 0.50. The receiver operating characteristic (ROC) curve is generated from several confusion matrices with cutoff values ranging from 0 to 1. Each unique cutoff lead to a certain sensitivity and 1-specificity that is plotted against each other to form the ROC curve. The area under the curve (AUC) provides a measure of model performance for any cutoff value and is the most widely used measure to compare binary classification models. Values closer to 1 are considered better classifiers than those close to 0.50, which suggest that a model is poor learner.

Prescriptive analytics

Prescriptive analytics refers to understanding what actions to take next. Once we identified a predictive model that best estimate the reasons for leaving, we formulate a decision (i.e. optimization) model that incorporates those estimated effects to help guide the decision maker in what decisions should be taken. This provides an HR professional an analytically-based means to decide what to do next to help improve employee retention overall. Example decisions could entail increasing employee pay, separating from poor performing employees whom are likely to leave anyway, or providing educational incentives that might reduce turnover. The practical usefulness is how to make these decisions across the entire workforce while accounting for all known constraints (e.g. salary budgets, job grade benefits). The decision model we develop provides a working example of this.

PREDICTIVE MODELS

We build and evaluate the models using different binary classification algorithms. Some of these methods have been suggested or used in the literature in estimating employee turnover. We investigate those approaches as well as other popular approaches used to support other business problems as shown in Table 3 previously.

Logistic Regression

Logistic Regression is the appropriate regression analysis to conduct when the dependent variable is a binary variable. Logistic regression is one the simplest and the most widely used of all classification techniques. While estimating the coefficients of each predictor (independent variables), logistic regression is somewhat similar to the linear regression except for the fact that the response is transformed using a link function known as a "logit", which assures that the outputs follow a logistic (sigmoid) curve. This assures that all predictions have values between 0 and 1, thus ensuring probabilities are generated as they defined. Maximum likelihood estimation is used to estimate the parameter coefficients of the model.

CART

Classification and Regression Trees (CART) is a modern, c.1984, are one of the most frequently used types of models used in business and scientific applications. Its advantage over other models is it is considered easy to interpret. In the classification tree setting, each predicted probability is derived from a set of IF-THEN rules provided by the decision tree. While trees tend to provide general understanding, they may not lead to the best predictive model performance. First, each terminal node (or leaf) in the tree is one possible probability prediction. This can lead to a small set of unique value predictions compared to other approaches such as logistic regression. Growing a tree so there is more splits can lead to more leafs, often leading to a larger set of unique probability predictions, but this often leads to overfitting the model. In our study, we tune each tree so as to optimize the tree complexity.

Random Forests

A Random Forest model consists of a collection or ensemble of decision trees, each capable of producing a response when presented with a set of predictor values. It just averages the probabilities generated by developing different decision trees. It has been shown ensembling many weak learning trees can improve overall performance. When ensembling trees one must consider how many trees to combine. This is considered a tuning parameter that one must find so as to not overfit to the training data.

Support Vector Machines

Support Vector Machines (SVM) are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects on the basis of different values for the categorical variable. It is appropriate when most of our predictors are numerical variables instead of categorical variables. In our study, we generated dummy variables for each categorical variable predictor before applying SVM. We found that the presence of some categorical predictors did not produce better results than other models used.

K-Nearest Neighbor

K-Nearest Neighbors algorithm (KNN) is a non-parametric method used for classification and regression. Both for the classification and regression of the input consists of the k-closest training examples in the feature space. The output is dependent on whether if we are using for the regression or classification. For regression-type problems, the output is the property value for the object. This value is the average of the values of its k nearest neighbors. For classification-type problems, the output is class membership. If the value of the k = 1, then the object is simply assigned to the class of that single nearest neighbor.

RESULTS

Descriptive Analysis:

The features available in our study were examined to identify any potential trends or causes to employee turnover. As stated in the motivation of our paper, the retail collaborator was concerned with understanding why employee turnover had increased above expectations for certain positions. Their goal was to understand why turnover occurred and be able to take the right action to retain their employees.

The wage a person is paid can affect if a person leaves or not. The plots in Figure 5 show the distribution of hourly pay ($) across all employees. The distribution among job types (e.g. A,B,C,D) were not statistically different. We derived a new variable called Pnormal that accounts for the market in which the position is located. This essentially serves as an index where 0 indicates that an employee is paid the average rate for that job in their particular market. Above 0 indicates an employee is paid above average, while below 0 indicates than an employee is paid below market rate.

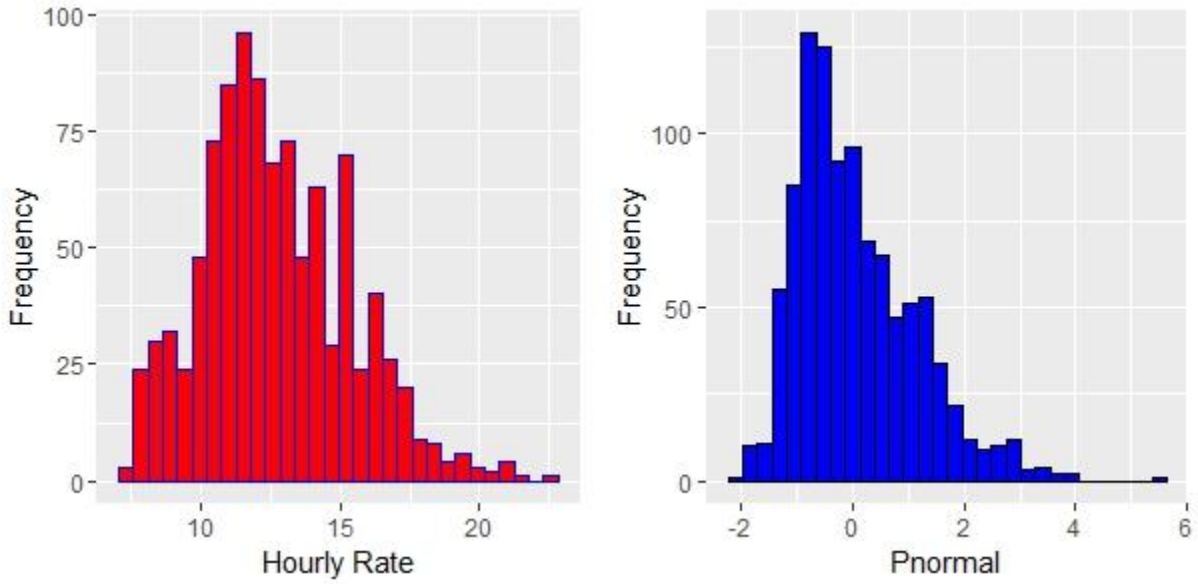

Figure 5: Distribution of hourly rate and normalized rate based on market (Pnormal)

Figure 6 provides the distribution of store performance based on the previous year to current year. We found that on average stores were performing slightly lower than the previous year based on

an organizational KPI. The spread of performance had more variance in the latest year, which might or might not be due to employee turnover. We do know that employee turnover in the current year was higher than in the previous year.

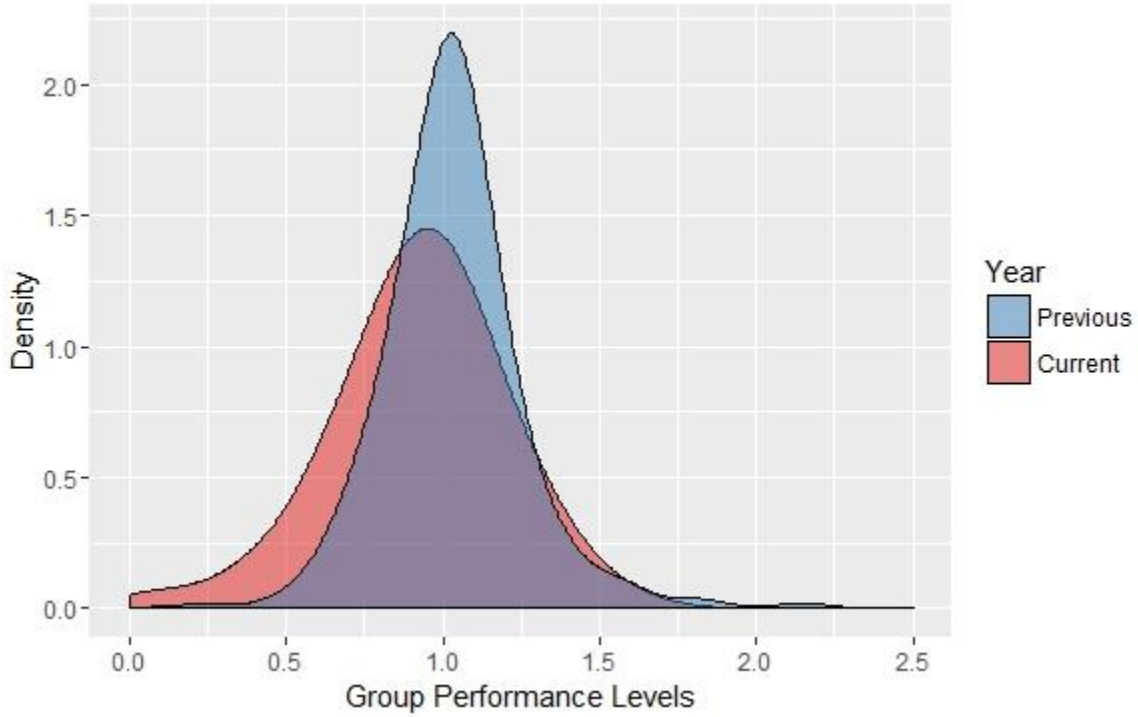

Figure 6: Group (store) performance year-on-year

The performance reviews of employees were provided based on an ordinal rank (i.e. Superior, good solid performer, needs improvement, and unsatisfactory). Those employees falling under needs improvement and unsatisfactory had HR policies in place to help improve their performance going further. These include further training and expectations over a specific future time window.

We did realize some interesting relationships in employee performance year-on-year as shown in Table 5. We found that 76% of those who moved from 'Superior' to 'Needs Improvement' left the firm. Also, the highest number of employees (187) were those who moved from 'Good Solid Performer' to 'Needs Improvement' and 46% of these employees left the firm. Overall from this table, it can be seen that the performance level of most of the employees that left the firm declined in current year as compared to the previous year. We discussed this with firm stakeholders and found that manager evaluation training for all stores was already being considered.

Employee Performance (Previous Year)	Employee Performance (Current Year)	Left	Still Employed	Total	Percent left
Superior	Superior	7	10	17	41.18%
Superior	Good Solid Performer	8	20	28	28.57%
Superior	Needs Improvement	19	6	25	76.00%
Superior	Unsatisfactory	12	6	18	66.67%
Good Solid Performer	Superior	15	43	58	25.86%
Good Solid Performer	Good Solid Performer	77	105	182	42.31%
Good Solid Performer	Needs Improvement	87	100	187	46.52%
Good Solid Performer	Unsatisfactory	37	20	57	64.91%
Needs Improvement	Superior	9	11	20	45.00%

Needs Improvement	Good Solid Performer	32	58	90	35.56%
Needs Improvement	Needs Improvement	105	75	180	58.33%
Needs Improvement	Unsatisfactory	60	28	88	68.18%
Unsatisfactory	Good Solid Performer	6	5	11	54.55%
Unsatisfactory	Needs Improvement	8	7	15	53.33%
Unsatisfactory	Unsatisfactory	17	7	24	70.83%

Table 5: Performance levels in previous year and current year

Figure 7 provides additional evidence that shows that good performers declined in the current year and needs improvement and unsatisfactory performance reviews increased.

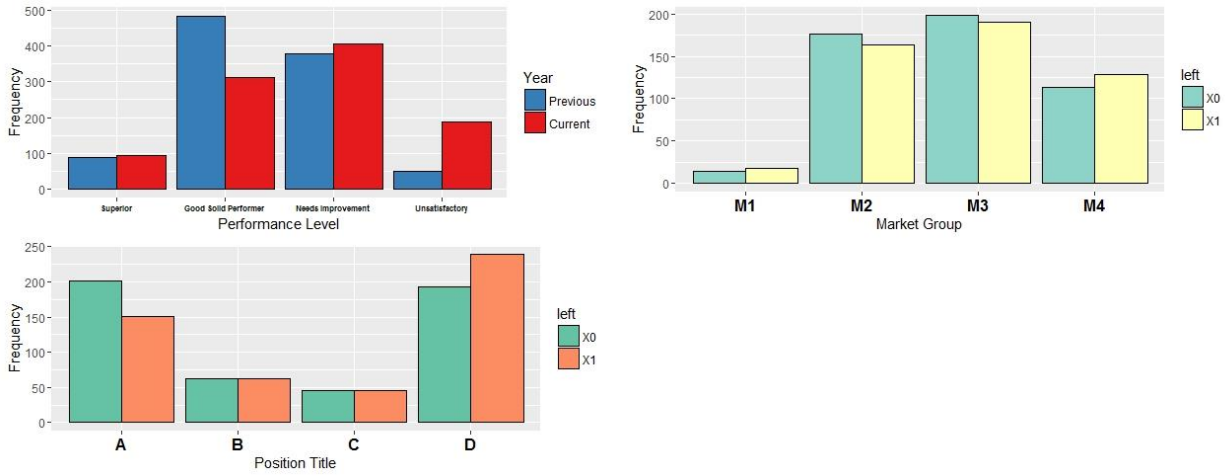

Figure 7: Examination of qualitative variables by those that left and stayed.

We explored the behavior of employees under the four job titles we considered. From Table 6 it can be clearly seen that the employees with the highest turnover rate are job D followed by job C, and then job B. Table 7 shows that 26% of the employees who resigned from the position of job A were above the salary range of their market class while only 2% of them were below salary range. The opposite trend can be seen in the case in job C and job D, where most of the employees who left were below the salary range than those above. Table 8 shows a similar overall trend across all four job titles. It can be seen that the average hourly rate of the employees left is even lower than the average hourly rate of overall employees.

Job Title	Employees left	Total Employees	%Employees left
A	151	352	43%
B	63	125	50%
C	46	91	51%
D	239	432	55%

Table 6: Turnover rate vs job title

Job Title	Employees left	%Employees above the salary range	%Employees below the salary range
A	151	26%	2%
B	63	5%	2%
C	46	2%	20%
D	239	1%	10%

Table 7: Employees outside of the salary range that left vs. job title

Job Title	Hourly Rate of employees left	Hourly Rate of overall employees
A	14.76	15.07

B	8.79	9.15
C	11.84	12.08
D	11.66	11.89

Table 8: Average Hourly rate of employees left compared to overall employees vs. job title

Predictive Analysis:

We build various predictive models and evaluated their performance. *Table 9* shows the statistical performance we were able to achieve on this data set for each model.

	Training				Testing			
Models	Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC
Random Forest	100.00%	100.00%	100.00%	1.00	64.88%	69.13%	60.67%	0.68
k-Nearest Neighbour	63.77%	61.43%	66.10%	0.70	60.54%	57.72%	63.33%	0.67
Decision Tree (CART)	65.34%	64.86%	65.81%	0.72	62.88%	67.79%	58.00%	0.64
Logistic Regression	65.91%	62.86%	68.95%	0.72	63.21%	63.76%	62.67%	0.63
Support Vector Machines	70.90%	76.29%	65.53%	0.71	60.87%	63.09%	58.67%	0.61

Table 9: Confusion matrix metrics for different models for Testing Dataset

The drivers we found to be important using each model. For the most part, each model was suggesting that the normalized market rate index (Pnormal), hourly wage, position title, the region where an employee works, and store performance were important predictors.

To compare models, we used an ROC curve shown in Figure 8. The curve that is closest to the point c(1,1). We found that the random forest led to the greatest AUC (0.68), but it was extremely overfit compared to the training set. The second best performing model was kNN having an AUC of 0.67. We decided to use the decision tree (i.e. CART) as the final model of choice. This model

was interpretable to the HR practitioners and allowed us to use tree splits in our decision model to follow.

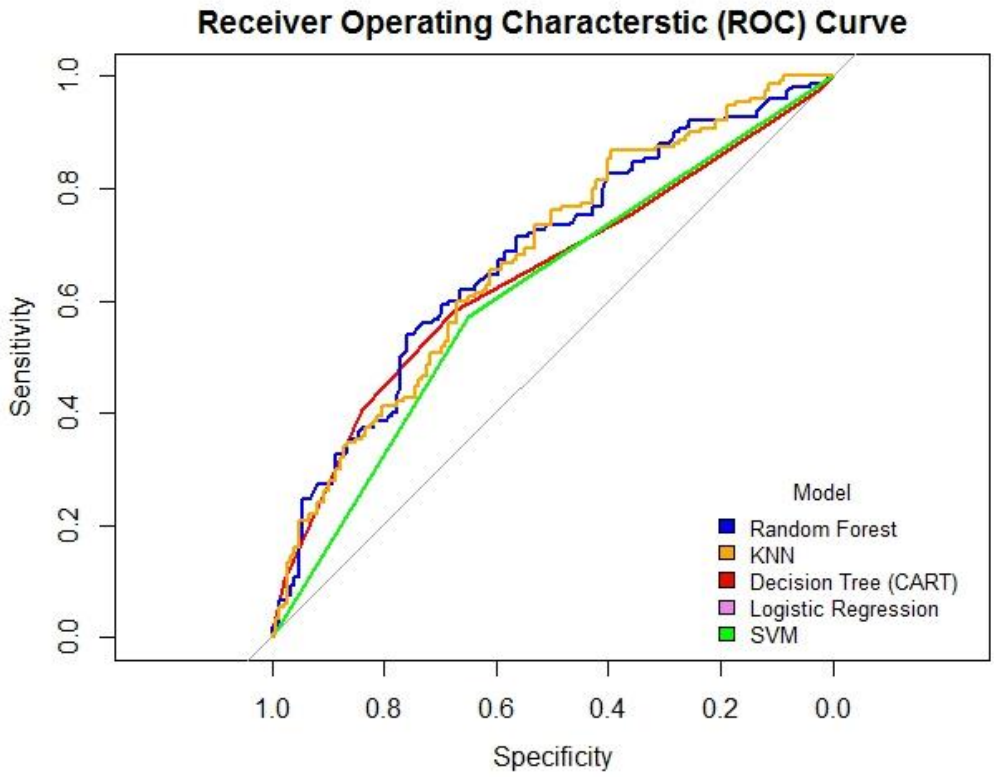

Figure 8: ROC curve showing the performance of each model

Figure 9 shows the decision tree created from CART. The performance of the store in the current year and previous year (GroupPerfCurrentYr, GroupPerfPrevYr,), hourly rate, region, and normalized wage index (Pnormal) were important drivers at explaining the probability that an employee will turn. We used some of these features (hourly rate, Pnormal) as decisions in our decision model. The location of where the employee works (region and store performance) is accounted for when such decisions are made.

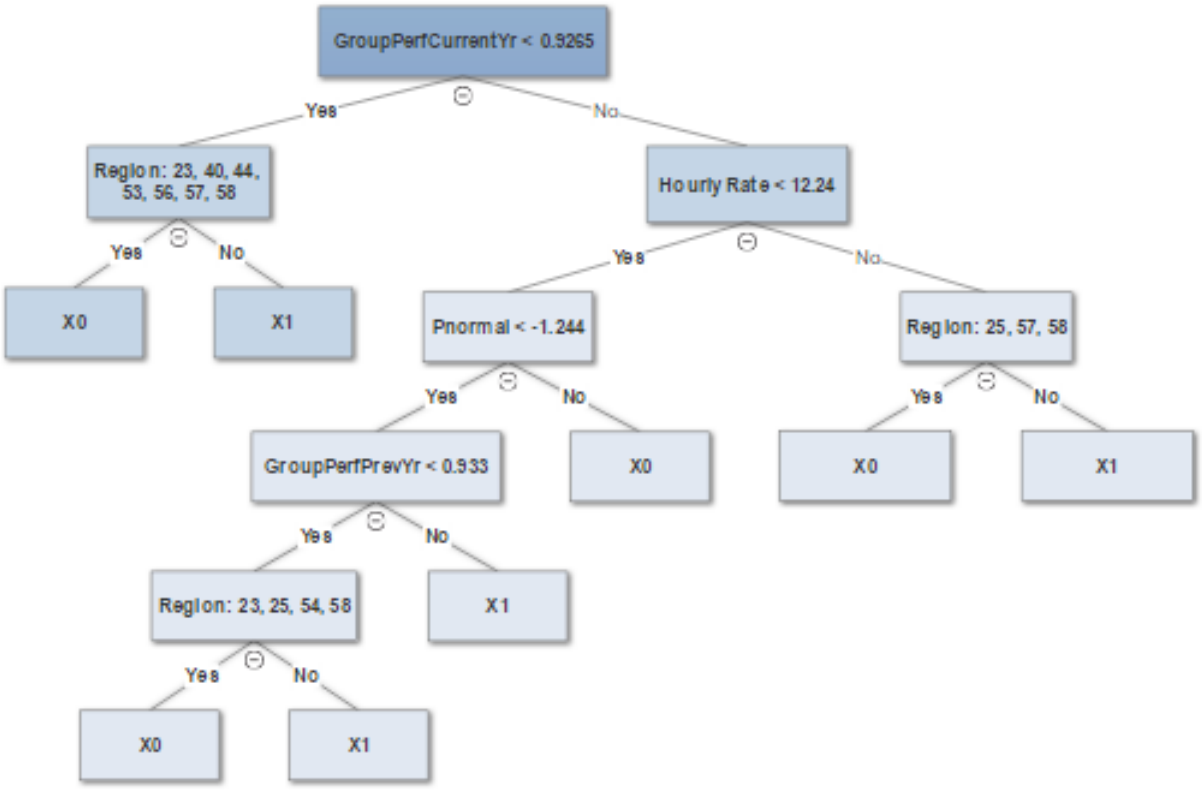

Figure 9: Decision tree (CART)

Prescriptive Analysis/Decision Model

Using the information from the decision tree we incorporate that information into a decision model to support HR decisions over the next planning horizon. The decision model describes the cause-effect relationship where controllable actions can be taken by decision-makers with the assumption that such decisions/actions will lead to an effect with respect to certain pre-defined performance measures. This is the first step in the prescriptive analysis. Next, the modeler will need to solve the problem using some optimization routine. The routine is behind the scope of this research, and is really not important. Practically we just want to know if the model can be solved, could it be solved in a reasonable amount of time so as to provide guidance to the decision-maker when they need it. We define the mathematical notation as follows:

Terms and definitions:

N = total number store employees

M = total number stores

A_{ij} = employee i in store j for job position A; $i = 1,..,A;\ j = 1,..,M$

B_{ij} = employee i in store j for job position B; $i = 1,..,B;\ j = 1,..,M$

C_{ij} = employee i in store j for job position C; $i = 1,..,C;\ j = 1,..,M$

D_{ij} = employee i in store j for job position D; $i = 1,..,D;\ j = 1,..,M$

e_{ij} = experience (years) of employee i in store j; $i = 1,..,N;\ j = 1,..,M$

τ_{ij} = latest performance review of employee i in store j; $i = 1,..,N;\ j = 1,..,M$

φ_{ij} = hourly rate ($) of employee i in store j; $i = 1,..,N;\ j = 1,..,M$

ω_{ij} = wage index of employee i in store j; $i = 1,..,N$; $j = 1,..,M$
ρ_{ij} = estimated probability of turn in next 6 months of employee i in store j; $i = 1,..,N$; $j = 1,..,M$
ψ_{ij} = estimated class of turn in next 6 months of employee i in store j; $i = 1,..,N$; $j = 1,..,M$;
$\psi_{ij} \in \{0,1\}$
φ^*_{ij} = **new** hourly rate (\$) of employee i in store j; $i = 1,..,N$; $j = 1,..,M$
ω^*_{ij} = **new** wage index of employee i in store j; $i = 1,..,N$; $j = 1,..,M$
ρ^*_{ij} = **new** estimated probability of turn in next 6 months of employee i in store j; $i = 1,..,N$; $j = 1,..,M$
ψ^*_{ij} = **new** estimated class of turn in next 6 months of employee i in store j; $i = 1,..,N$; $j = 1,..,M$;
$\psi^*_{ij} \in \{0,1\}$
K_j = the average team performance of job A at store location j; $j = 1,..,M$
Z_j = the average team performance of job B at store location j; $j = 1,..,M$
Y_j = the average team performance of job C at store location j; $j = 1,..,M$
U_j = the average team performance of job D at store location j; $j = 1,..,M$
A = the next sixth month budget (\$) for Job A
B = the next sixth month budget (\$) for Job B
C = the next sixth month budget (\$) for Job C
D = the next sixth month budget (\$) for Job D

We only consider two potential decisions. First, how much should HR increase the wage of a specific employee given their job type and market. Second, should certain employees be terminated immediately because they are already predicted with high probability to leave the company anyway.

Decision variables
x_{ij} = amount to increase wage of employee i in store j; $i = 1,..,N$; $j = 1,..,M$
y_{ij} = decision to fire employee i in store j, $y_{ij} \in \{0,1\}$; $i = 1,..,N$; $j = 1,..,M$

Our objective is to maximize the percentage of expected non-turners so as to have a workforce that can complete the workload over the next planning horizon. Today is we assume that we make no changes, a retailer might estimate that 65% of their workforce will be available in the future. However, if the retailer takes action (i.e. changes their decision variables), this will reduce the probability that certain employees will leave during the next window, which should yield a higher expected workforce percentage.

Objective function:
$max[\sum_j \sum_i \psi^*_{ij}]/(N * M)$ (maximize the percentage of expected non-turners to complete workload)

Constraints:
$\sum_i \tau_{ij}/A \geq K_j \quad \forall j$ (average job A performance should exceed some threshold)
$\sum_i \tau_{ij}/B \geq Z_j \quad \forall j$ (average job B performance should exceed some threshold)
$\sum_i \tau_{ij}/C \geq Y_j \quad \forall j$ (average job C performance should exceed some threshold)
$\sum_i \tau_{ij}/D \geq U_j \quad \forall j$ (average job D performance should exceed some threshold)
$\sum_j \sum_i A_{ij} \leq \mathbf{A}$ (budget for job-type A must be satisfied)
$\sum_j \sum_i B_{ij} \leq \mathbf{B}$ (budget for job-type B must be satisfied)

$\sum_j \sum_i C_{ij} \leq \mathbf{C}$ (budget for job-type C must be satisfied)
$\sum_j \sum_i D_{ij} \leq \mathbf{D}$ (budget for job-type D must be satisfied)
$\sum_j \sum_i \omega_{ij}^* / N \geq 0.95$ (average wage index of job A is at least 0.95; 1 would imply market avg.)
$x_{ij} \geq 0$ (hourly wages can only increase)
$y_{ij} \in \{0,1\}$ (it is worth the risk to fire some employees now because they are poor performers and have a high probability of leaving anyways in next six months)

CONCLUSIONS & FUTURE RESEARCH

Employee turnover is a major workforce management challenge for retailers. The retail domain has higher turnover than other industries which makes it especially important to be able to manage efficiently. The rise in data analytics has seen remarkable feats in retailer's ability to understand their customers. Unfortunately, the same investment has not been realized in better understanding your employees so as to improve a retailer's workforce.

The objective of this study was to develop an analytically-based framework that HR professionals in retail and potentially other domains (e.g. manufacturing) can use to effectively use their data to better management their workforce. We provide this by using data from a regional retailer and show how interfacing descriptive, predictive, and prescriptive analytics can provide a means to identify cause-effect relationships, predict what will happen in the future, and then how to take action so as to maximize your workforce (or minimize your turnover).

We plan to continue to develop our decision model to account for other items behavioural scientists have found important in explaining employee turnover. In our study, we could only identify potential causes and estimate effects among a limited set of features. However, we know HR areas could measure more about their employees to gain better insight about them. We posit that providing engagement survey's over time and incorporating such information into the predictive models could reveal additional causes of turnover that are not captured in our study.

REFERENCES

Ashford, S. J., et al. (1989). "Content, cause, and consequences of job insecurity: A theory-based measure and substantive test." Academy of Management journal **32**(4): 803-829.

Barrick, M. R. and R. D. Zimmerman (2005). "Reducing voluntary, avoidable turnover through selection." Journal of applied psychology **90**(1): 159.

Brust, A. (2013) Five Big Data Trends Revolutionizing Retail.

> More retailers are finding that Big Data can revitalize an industry challenged by a slow economy, increasingly empowered consumers, mobile proliferation and an ever-growing number of channels.

Cadez, I. V. and P. Smyth (2001). Bayesian Predictive Profiles With Applications to Retail Transaction Data. NIPS.

Collins, C. J. and K. D. Clark (2003). "Strategic human resource practices, top management team social networks, and firm performance: The role of human resource practices in creating organizational competitive advantage." Academy of Management journal **46**(6): 740-751.

Dalton, D. R. and W. D. Todor (1979). "Turnover turned over: An expanded and positive perspective." Academy of management review **4**(2): 225-235.

Dunne, P., et al. (2013). Retailing, Cengage Learning.

Fatima, A. and S. Rahaman (2014). "Mining System in HR: A Proposed Model." International Journal of Computer and Information Technology **3**(05): 2279-0764.

Feffer, M. (2014). "HR Moves toward Wider Use of Predictive Analytics." Society for Human Resource Management.

Fidalgo, F. and L. B. Gouveia (2012). "Employee turnover impact in organizational knowledge management: The Portuguese real estate case." Journal of Knowledge Management, Economics and Information Technology **2**(2): 1-16.

Fulmer, I. S., et al. (2003). "Are the 100 best better? An empirical investigation of the relationship between being a "great place to work" and firm performance." Personnel Psychology **56**(4): 965-993.

Graen, G. B., et al. (1982). "Role of leadership in the employee withdrawal process." Journal of applied psychology **67**(6): 868.

Handfield-Jones, H., et al. (2001). "Talent management: A critical part of every leader's job." Ivey Business Journal **66**(2): 53-74.

Hatch and Dyer (2004). "Human capital & leasing as a resource of sustainable competitive advantage." Strategic Management Journal.

Hoffman, D. L. and M. Fodor (2010). "Can you measure the ROI of your social media marketing?" MIT Sloan Management Review **52**(1): 41.

Holtom, B. C., et al. (2008). "5 Turnover and Retention Research: A Glance at the Past, a Closer Review of the Present, and a Venture into the Future." Academy of Management annals **2**(1): 231-274.

Hom, P. W., et al. (2008). "Challenging conventional wisdom about who quits: revelations from corporate America." Journal of applied psychology **93**(1): 1.

Hübner, A. H. and H. Kuhn (2012). "Retail category management: State-of-the-art review of quantitative research and software applications in assortment and shelf space management." Omega **40**(2): 199-209.

Huselid, M. A. and B. E. Becker (2006). "Improving Human Resources' analytical literacy: lessons." THE FUTURE OF HUMAN RESOURCE MANGEMENT: 278.

Jackson, S. E., et al. (1986). "Toward an understanding of the burnout phenomenon." Journal of applied psychology **71**(4): 630.

Kacmar, K. M., et al. (2006). "Sure everyone can be replaced… but at what cost? Turnover as a predictor of unit-level performance." Academy of Management journal **49**(1): 133-144.

Lee, C., et al. (1990). "Interactive effects of "Type A" behavior and perceived control on worker performance, job satisfaction, and somatic complaints." Academy of Management journal **33**(4): 870-881.

Lee, T. W. and T. R. Mitchell (1994). "An alternative approach: The unfolding model of voluntary employee turnover." Academy of management review **19**(1): 51-89.

McPherson, J. M., et al. (1992). "Social networks and organizational dynamics." American sociological review: 153-170.

Michaels, E., et al. (2001). The war for talent, Harvard Business Press.

Miller, T. (2009) Freedom Is Still the Winning Formula.

Mirvis, P. H. and E. E. Lawler (1977). "Measuring the financial impact of employee attitudes." Journal of applied psychology **62**(1): 1-8.

Mishra, S. N. and D. R. Lama (2016). "A Decision Making Model for Human Resource Management in Organizations using Data Mining and Predictive Analytics." International Journal of Computer Science and Information Security **14**(5): 217.

Mishra, S. N., et al. (2016). "Human Resource Predictive Analytics (HRPA) for HR Management in Organizations." International Journal of Scientific & Technology Research **5**(05): 33-35.

Mladenic, D., et al. (2001). "Exploratory analysis of retail sales of billions of items." Computing Science and Statistics **33**.

Mobley, W. H. (1982). "Some unanswered questions in turnover and withdrawal research." Academy of management review **7**(1): 111-116.

Mobley, W. H. (1992). "Employee turnover: Causes, consequences, and control."

O'Reilly, C. A., et al. (1991). "People and organizational culture: A profile comparison approach to assessing person-organization fit." Academy of Management journal **34**(3): 487-516.

Payne, S. C. and A. H. Huffman (2005). "A longitudinal examination of the influence of mentoring on organizational commitment and turnover." Academy of Management journal **48**(1): 158-168.

Pfeffer, J. (1985). "Organizational demography: Implications for management." California Management Review **28**(1): 67-81.

Pfeffer, J. and A. Davis-Blake (1992). "Salary dispersion, location in the salary distribution, and turnover among college administrators." Industrial & labor relations review **45**(4): 753-763.

Price, J. L. (1977). The study of turnover, Iowa State Press.

Sadath, L. (2013). "Data Mining: A Tool for Knowledge Management in Human Resource." International Journal of Innovative Technology and Exploring Engineering **2**(6): 2278-3075.

SEBT, M. V. and H. YOUSEFI (2015). "Comparing data mining approach and regression method in determining factors affecting the selection of human resources." Cumhuriyet University Faculty of Science Journal **36**(4): 1846-1859.

Shashanka, M. and M. Giering (2009). Mining Retail Transaction Data for Targeting Customers with Headroom-A Case Study. Artificial Intelligence Applications and Innovations III, Springer**:** 347-355.

Staw, B. M. (1980). "The consequences of turnover." Journal of occupational Behaviour: 253-273.

Viator, R. E. and T. A. Scandura (1991). "A study of mentor-protégé relationships in large public accounting firms." Accounting Horizons **5**(3): 20.