Disney

Retention Risk &
Employee Turnover Strategy
Initial Findings

Prepared for Dr. Donald Wedding, CHRO
Annual HR Summit Day 2
May 12, 2019

People Insights Team
Agnes Cheng
Brent Young
Mary Taylor
Michelle Eure

Disclaimer*: Please note that the data is from Kaggle and is fictional. This project was completed as part of the MSDS 498 Capstone Project course within the Northwestern University - Data Science Program. All data, dashboards, and insights used throughout this project are completely simulated and not in any way connected to or a reflection of The Walt Disney Company. Please do not duplicate or distribute outside of the context of this course.*

Problem Statement

Like many organizations around the world, The Walt Disney Company has undergone significant organizational and leadership changes over the last 5 years so that they can remain competitive in the marketplace. As a result, this creates a feeling of constant change, uncertainty, instability, and an increase in turnover.

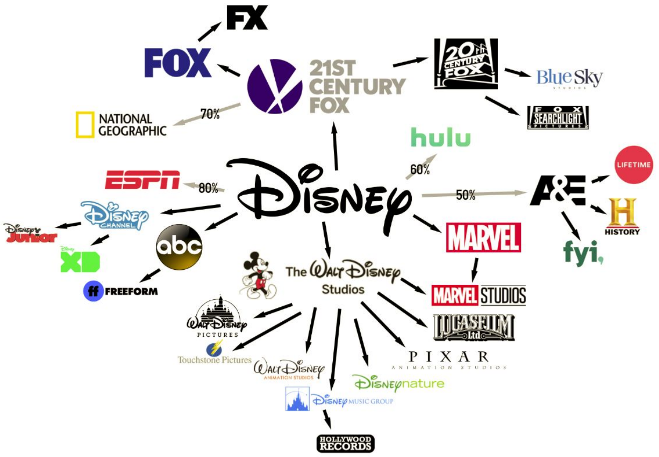
Turnover has increased.

Strong competition for talent, especially in technology, finance, legal from local competitors.

Need to retain key employees amidst significant org changes and Disney/21CF integration.

Clients' 'hunger' to use data driven approaches to solve people issues.

To add to this, in mid-March 2019, The Walt Disney Company acquired 20th Century Fox. With the recent acquisition, leaders and HR executives anticipate even more disruption, which may lead to the possible loss of high potential and key talent. Additionally, competition for talent, specifically within functions such as Finance, Technology, and Legal continue to remain fierce from local entertainment and technology competitors. This is primarily due to the fact that "content" is king and more and more companies are creating original programming and launching direct-to-consumer entertainment (e.g., SVoD service) and need the talent to support their business. Lastly, Corporate HR has received feedback from leaders within the business that they would love to see an increased use of employee data to help them make data-driven decisions to solve people issues in regards to their talent.

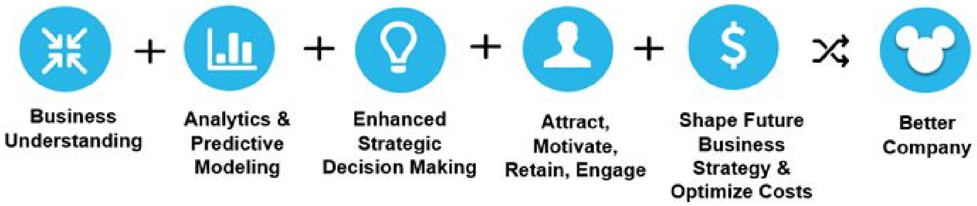


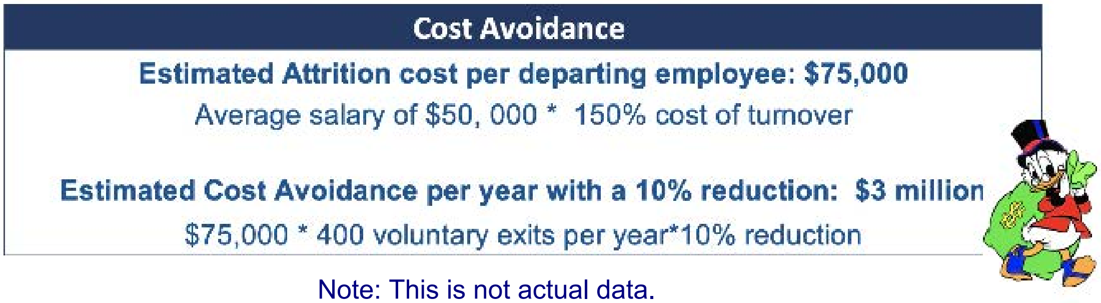
Repercussions of High Voluntary Turnover

Here are some repercussions of high voluntary turnover:

- High replacement costs for the company.
- Projects get delayed in a time where meeting deadlines are crucial.
- Work needs to be redistributed to the remaining employees, which causes worker fatigue, burnout, and employee dissatisfaction.
- New talent has to be recruited, trained and given time to acclimatize themselves to the company.
- Employee satisfaction goes down**.**
- Loss of institutional knowledge.

To help Disney Corporate overcome these challenges, the People Insights' team will utilize a data set that contains approximately 1,500 employee records to build several classification models to predict the probability that an employee will leave the company, determine which factors prompt employees to stay vs. leave, and create a Stay Survey so that leaders/HR can use the model results to retain high-risk employees. Overall, this will help inform the development of a Corporate HR strategy to retain top talent, increase employee satisfaction/engagement survey scores, increase productivity, decrease voluntary turnover, and lower hiring/replacement costs, thus saving the company money. An interactive dashboard and mobile interface to share the analysis with Corporate HR and leaders was also created.


Business Understanding
+
Analytics & Predictive Modeling
+
Enhanced Strategic Decision Making
+
Attract, Motivate, Retain, Engage
+
Shape Future Business Strategy & Optimize Costs
Better Company



Cost Avoidance

Estimated Attrition cost per departing employee: $75,000

Average salary of $50, 000 * 150% cost of turnover

Estimated Cost Avoidance per year with a 10% reduction: $3 million

$75,000 * 400 voluntary exits per year*10% reduction

Note: This is not actual data.

Description of the Data

The People Insights team will use fictional data from the IBM HR Analytics Employee Attrition data set (Kaggle), which contains 1470 records and 35 employee variables that can be broken down into 4 categories: demographic, employee job/performance, employee history, and employee survey. Attrition represents our classification response variable in regards to whether the employee left the company (0 = No, 1 = Yes). There is a total of 237 people who voluntarily left the company and 1233 people who stayed in the data set. This equates to a voluntary turnover rate of 16%. As a result, the data set is clearly imbalanced.

Demographic variables include age, education, education field, gender, marital status, and over 18 (Y/N). There are also several variables related to the employee's job and performance at the company such as business travel (Y/N), salary, department, distance from home to work, job level, job role, overtime eligibility, percent salary increase, performance rating, standard work hours, and stock option level. Furthermore, the data set includes historical information on the employee such as the number of companies worked, total working years, tenure, training time, years in current role, years since last promotion, and years with current manager. Lastly, employee survey results pertaining to environment satisfaction, job involvement, job satisfaction, relationship satisfaction, and work-life balance are also provided. The following table provides a high-level summary and description of the full data set.

Category	Description	Additional Information
File Format	Comma delimited (CSV) file	Header rows
Number of Records	1470	Imbalanced data set
Number of Fields	35	Demographic Employee job/performance Employee history Employee survey

Response Variable	Whether the employee left the company (0 = No, 1 = Yes)	Yes = 237
		No = 1233

Variable Types	Mix of numeric, categorical, and binary.	
Number of Missing Records	0	No imputation was conducted
Time Frame	2015 - 2016	

Please refer to Appendix 1 for the data dictionary.

The data was split into a training (70%) and test set (30%). The training set has 1029 records and the test set has 441 records. The training set will be used to fit the models and the test set will be used to estimate prediction error for model selection. K-fold cross-validation, which is a resampling procedure, was also conducted on the training dataset given the limited number of employee records in the data set and was used for model selection purposes. K-folds cross validation also allows us to estimate how the model is expected to perform on unseen data. The best model will then be applied to the test set for final evaluation.

Overview of the Data

Prior to modeling, the People Insights team conducted exploratory data analysis (EDA) on the full data set so that we could get a better understanding of the data, find initial trends, understand the level of data cleaning necessary, and develop insights that will help inform which variables to include in our models. The analysis was completed using Python and a variety of visualization tools such as Matplotlib and Seaborn. The code was shared using Google's CoLab. During this stage, we utilized histograms, barplots, and descriptive statistics to get a better understanding of the distributions of each variable. See appendix 2 for additional details.

Here are some general observations:

- The data for employees primarily comes from the Research & Development and Sales departments, with Human Resources having the least amount of employees. Due to those specific department segments, most of the nine job roles are primarily Sales Executives, Research Scientists, and Laboratory Technicians. With this limitation, we would need to keep in mind that applying our model to other departments or other job roles may not be as accurate. In future scenarios, if we decide to apply this model throughout all departments or additional departments, then we may need to retrain our model for accuracy.
- The data set is imbalanced with a low rate of attrition *(237 out of 1470 employees left the company)* so we are naturally trying to predict "rare events".
- Majority of the variables were skewed such as Distance from Home, Job Level, Monthly Income, Number of Companies Worked, etc.

We also analyzed predictor relationships with the attrition response variable using stacked bar plots for categorical variables and notched boxplots for numeric variables to see if there was high variability and whether or not the median differences between whether a person stayed or left the company is wide. This helped inform which variables we selected for our models. See appendix 3 & 4 for visuals.

The following **numeric** variables below had the most variability and median differences compared to attrition. This indicates that these variables are potentially strong predictors to include in our models. *Note: The italicized variables below were included in our models.*

- *Age, Daily Rate, Distance from Home*, Job Level, Monthly Income, *Number of Companies Worked*, Stock Option Level, *Total Working Years, Training Times Last Year, Years at the Company,* Years in Current Role, Years with Current Manager. Note: Survey related variables did not show a median difference, but when using the mean, variables such as *job satisfaction* and *work environment satisfaction***,** had mean differences that were wide.

The following **categorical** variables had the most differences compared to attrition.

- Business Travel, *Marital Status*, Education Field, *Overtime*

A correlation matrix (see appendix 5) also helped us determine if some of the predictors are highly correlated with each other so that only one of those variables are included in the automated feature selection techniques. Using this approach also removed a variable from each pair of collinear variables so that automated feature selection techniques didn’t end up choosing similar and collinear variables. Removing similar variables also helped eliminate multicollinearity concerns. The correlation matrix revealed several variables that had correlations greater than 0.70:

- Percent Salary Hike vs. Performance Rating (0.77)
- Years At Company vs. Years With Current Manager (0.77)
- Years At Company vs. Years In Current Role (0.76)
- Monthly Income vs. Job Level (0.89)
- Monthly Income vs. Total Years Worked (0.70).

As a result, Percent Salary Hike, Years With Current Manager, Years In Current Role, Job Level, and Monthly Income were removed accordingly. We also dropped variables that contained the same value for all employees such as Employee Count, Over 18, and Standard Hours.

Feature Selection: The second part of EDA, focused on using automated variable selection techniques such as univariate feature selection, recursive feature elimination (with logistic regression), tree-based feature selection (feature importance), a decision tree, and lasso embedded method on the training data set. A wide variety of features were identified across each of the options evaluated.

The following variables below appeared the most frequently (see appendix 6) and/or had the highest importance using the variable selection approaches. This indicates that these variables are potentially strong predictors to include in our models:

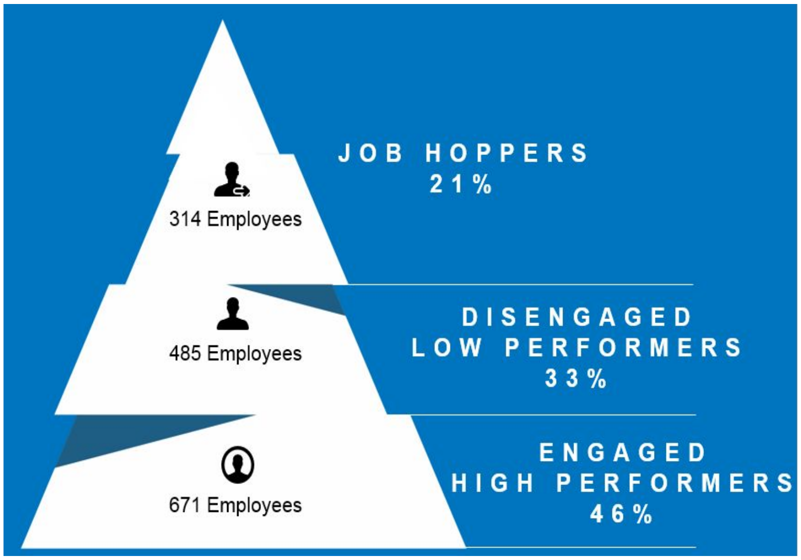
- Age, Daily Rate, Years at the Company, Total Working Years, Overtime, Distance from Home, Marital Status, Job Satisfaction, Environment Satisfaction, Number of Companies Worked.

Overall, here is a list of the features that we decided to move forward with based on what we saw in our EDA and our automated variable selection techniques:

- Age, Daily Rate, Distance from Home, Number of Companies Worked, Total Working Years, Training Times Last Year, Years at the Company, Job Satisfaction, Work Environment Satisfaction, Marital Status, Overtime, Relationship Satisfaction, Job Involvement, and Work-Life Balance. *Note: Performance Rating was not included in the model since Disney has recently moved to "no ratings". As a result, this variable will no longer be available in the future, but "shadow ratings" will be available.*

Cluster Analysis: The third part of EDA, focused on using cluster analysis (e.g., k-means clustering) to develop clusters using the survey and demographic data so that we can better understand how employees are feeling. This allows HR to develop targeted strategies and policies to increase employee engagement and satisfaction, which can ultimately improve employee retention.

Here are 3 segments that we determined based on the cluster analysis using demographic proportions and survey averages within each cluster. We've also profiled each segment as well. R was used for this analysis. See appendix 7 for the actual clusters.



Segment 1: “Job Hoppers” (314; 21% of Respondents):

This group dislikes the company’s work environment and are indifferent about their satisfaction with their job. Historically, they don’t stay at companies very long because they are money and promotion chasers. As a result, they tend to make a lot of money and many of them have worked for a number of companies in the past. They are high performers, are highly involved in their work, and have great relationships with their colleagues. Some tend to have long commutes, which may explain why their work-life balance and work environment scores are the lowest amongst their peers. Interestingly, **26%** of the people in this group left the company (80 out of 314 employees), the highest voluntary turnover rate compared to the other segments.

Segment 2: “Disengaged Low Performers” (485; 33% of Respondents):

This group loves the company’s work environment, but are somewhat dissatisfied/indifferent about their job. They receive lower performance ratings, lower annual salary increases, lower income, and the least training and development amongst their peers. They tend to have short commutes, which may explain why their work-life balance and work environment scores are the highest amongst their peers. They have the least involvement in their work amongst their peers, but they have good relationships with their colleagues. **16%** of the people in this group left the company (76 out of 485 employees).

Segment 3: “Engaged High Performers” (671; 46% of Respondents):

This group likes the company’s work environment and absolutely loves their job. They are highly committed to their role and to the company. As a result, many of them have been in their role and have worked for the company for a long time. They seem to be comfortable in their role and therefore aren’t constantly looking for promotional opportunities. These employees are also high performers and as a result, they are rewarded with the best performance ratings, highest salary increases, and stock options compared to their peers. They have a good monthly income and tend to have short commutes, which may explain their high work-life balance and work environment scores. Their relationships with their colleagues are good, but not great. As a result, this may slightly negatively impact their work environment satisfaction. Interestingly, **12%** of the people in this group left the company (81 out of 671 employees), the lowest voluntary turnover rate compared to the other segments.

Transformations

Using boxplots in Python, we conducted a thorough analysis of all the variables to detect potential outliers. After our review, we uncovered that the majority of the variables did not include any major outliers, except for monthly income. However, we decided not to handle/truncate the outliers in this particular variable because we did not use this variable in our models. Using summary statistics, we also checked for missing values. However, after careful inspection, the data set did not contain any missing values, so no missing value imputation was conducted. In regards to categorical data, our Chief Data Scientists often applied label encoding to variables such as Attrition, which was changed from Yes and No to 1 = Yes and 0 = No (e.g., numeric values). Further examples of this encoding can be seen in our data dictionary (see appendix 1). The People Insights team plans to create new variables in the next phase of the modeling which will help generate additional insights and potentially increase model performance. For example, we plan on exploring the use of the cluster segments that were derived from our cluster analysis and binning some of the numeric variables to create new features such as age range, distance to work groupings, etc. However, it’s important to note that if a cluster variable is used in future models, the People Insights team will need to create a model to predict cluster segment classifications that can then be added to the model. If not, the cluster variable cannot be used.

Analysis of Data

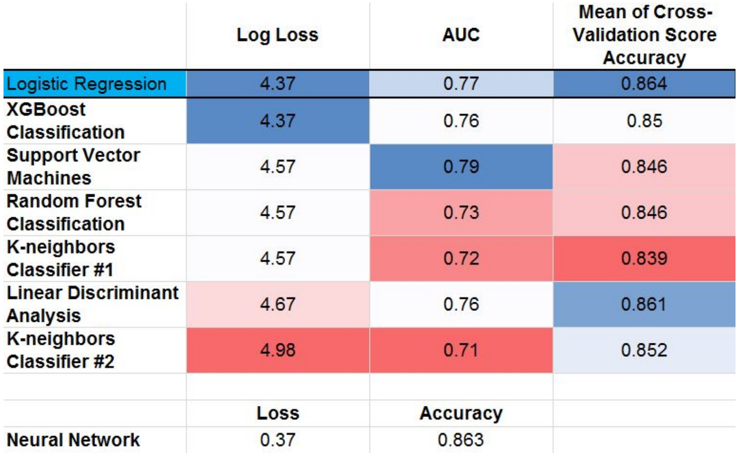
Given that this a classification problem, a mix of modeling methods were used; from explainable modeling techniques such as logistic regression and decision trees to more complex modeling techniques such as linear discriminant analysis, XGBoost, SVM, random forest, K-Nearest Neighbors and Neural Network. All modeling was completed in Python and the code was shared using CoLab.

The models were assessed using statistical evaluation criteria such as AUC, ROC Curves, Logarithmic loss (Logloss), mean k-folds cross-validation accuracy scores, and Decile Analysis. The ROC Curve measures how well a model can differentiate between True Positives (TP) and False Positives (FP), while Area Under the Curve or “AUC” is the percentage of the area that is under the curve (higher the AUC the better). The ideal ROC curve will show a bow shape, which means that the model does a good job on rank ordering the individual probabilities. See appendix 8 for additional model result details.

Logloss measures the performance of a classification model where the prediction input is a probability value between 0 and 1. The goal of our machine learning models is to minimize this value. As a result, a perfect model would have a Logloss of 0. Logloss increases as the predicted probability moves further and further from the actual label. For example, predicting a probability of .01 when the actual observation label is 1 would be bad and result in a high log loss. See appendix 8 for additional model result details.

Mean k-folds cross-validation accuracy scores were also incorporated to estimate how the model is expected to perform on unseen data. Additionally, given that the data set is imbalanced, confusion matrix metrics such as sensitivity and specificity were also produced and analyzed since accuracy can be a misleading metric when the data set is imbalanced (e.g., predicting rare events). See appendix 8 for additional model result details. Overall, the Logistic Regression model yielded the best performance metrics (see heat map below), which was great news considering it is the most interpretable, simple, and easy to explain. As a result, we decided to move forward with the Logistic Regression model.

Here is a heat map and summary of the model results:



	Log Loss	AUC	Mean of Cross-Validation Score Accuracy
Logistic Regression	4.37	0.77	0.864
XGBoost Classification	4.37	0.76	0.85
Support Vector Machines	4.57	0.79	0.846
Random Forest Classification	4.57	0.73	0.846
K-neighbors Classifier #1	4.57	0.72	0.839
Linear Discriminant Analysis	4.67	0.76	0.861
K-neighbors Classifier #2	4.98	0.71	0.852
	Loss	Accuracy	
Neural Network	0.37	0.863	

Determining Probability Cutoffs for Low, Medium, and High Risk

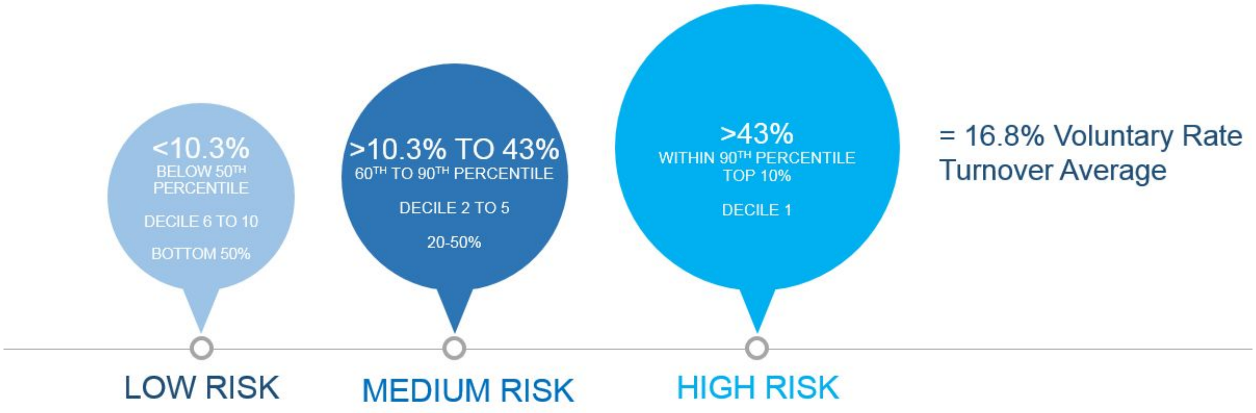
A decile analysis is a break out of ten different groups based on employee probabilities of leaving the company, with the first decile consisting of the group with the highest probabilities of leaving the company and the tenth decile consisting of the group with the lowest probabilities of leaving the company. A decile analysis is created to test the model's ability to predict the intended outcome. Each column in the decile analysis represents a collection of records that have been scored using the model. The height of each column represents the average of those records' actual attrition flag (where 0 = stayed with the company, 1 = left the company). When looking at a decile analysis, a "staircase effect" is the intended outcome, which means that we are binning the employees correctly. In other words, a decile analysis is basically a visual representation of the Logloss metric.

As a result, in order to help determine probability cutoffs for high, medium, and low risk employees, the People Insights team decided to use the decile analysis on the test data set for the logistic regression model. For instance, high risk individuals are employees who had a probability greater than 43%. These are individuals that fell in decile 1 (within 90th percentile or top 10%). Medium risk individuals are employees who had a probability between 10.3% to 43%. These are individuals that fell in deciles 2 to 5 (60th to 90th percentile or top 20-50%). Lastly, low risk individuals are employees who had a probability less than 10.3%. These are individuals that fell in deciles 6 to 10 (below 50th percentile or bottom 50%). The average probability for the test data set is 16.8%, which is close to the voluntary turnover rate of 16% on the full data set.

Here is a visual summary of the retention risk probability range based on the decile analysis:

Retention Risk Probability Range Based on Decile Analysis

Probability of leaving relative to the population.
Results on from the test data set.



<10.3%
BELOW 50TH PERCENTILE
DECILE 6 TO 10
BOTTOM 50%

>10.3% TO 43%
60TH TO 90TH PERCENTILE
DECILE 2 TO 5
20-50%

>43%
WITHIN 90TH PERCENTILE
TOP 10%
DECILE 1

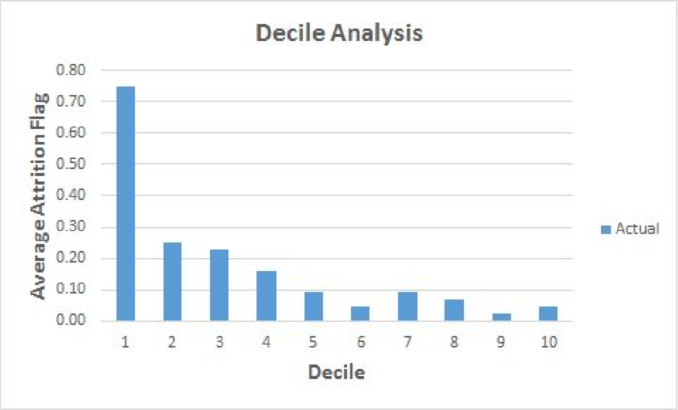
= 16.8% Voluntary Rate Turnover Average

LOW RISK

MEDIUM RISK

HIGH RISK

When looking at the decile analysis below, a “staircase effect” is shown below, which indicates that the model is doing a good job predicting the intended outcome and is binning the employees correctly. In fact, the top decile has an average attrition flag of 0.75. This is 3 times the 0.25 of the next decile. The last decile has an average attrition flag of 0.04.



Preliminary Conclusions

Given that our best performing model was Logistic Regression, we decided to use the model coefficients and odds ratios to identify the factors and themes that prompt employees to stay vs. leave the company so that the company can make changes to the work environment, policies, HR strategy, etc. This, in turn, will decrease voluntary turnover rate, decrease hiring/replacement costs, and increase employee satisfaction scores. Overall, when interpreting the model, everything made intuitive sense and confirmed what we saw in our EDA (e.g., box plots, bar plots) and cluster analysis results/profiles.

Our research on the use of analytics in human resources also identified data that could have been useful, but was not included in the Kaggle data set and/or is not available at this time. This includes work location state, home location state, zip codes, management level, people manager (Y/N), number of dependents, dependents (Y/N), compa ratio (pay compared to the market), immediate manager demographics, pension (Y/N), 401k (Y/N), pass usage (number of times Disney employee parks pass was used over the last year), distance from home to Disneyland, education reimbursement (Y/N), home owner (Y/N), work life flexibility (number of security buzz-ins into the office/parking lot over the last year), org change indicator (number of direct leader changes over the last year), and number of voluntary exits in the employee's immediate team. We advise Disney Corporate to reach out to the People Insights Data Operations team for further assistance on capturing this data in the future so that it can be included in future models. It might also be helpful to integrate additional employee survey related data, which can help increase model performance. Overall, the work and analysis that has been conducted has raised awareness and enhanced understanding of employee attrition at Disney.

Here are some general observations and learnings from the logistic regression model and why it makes intuitive sense. Total Working Years, Overtime (Y/N), Marital Status, and Distance from Home were the top 4 most predictive variables given the high coefficients compared to the others. The below variables are in order of importance.

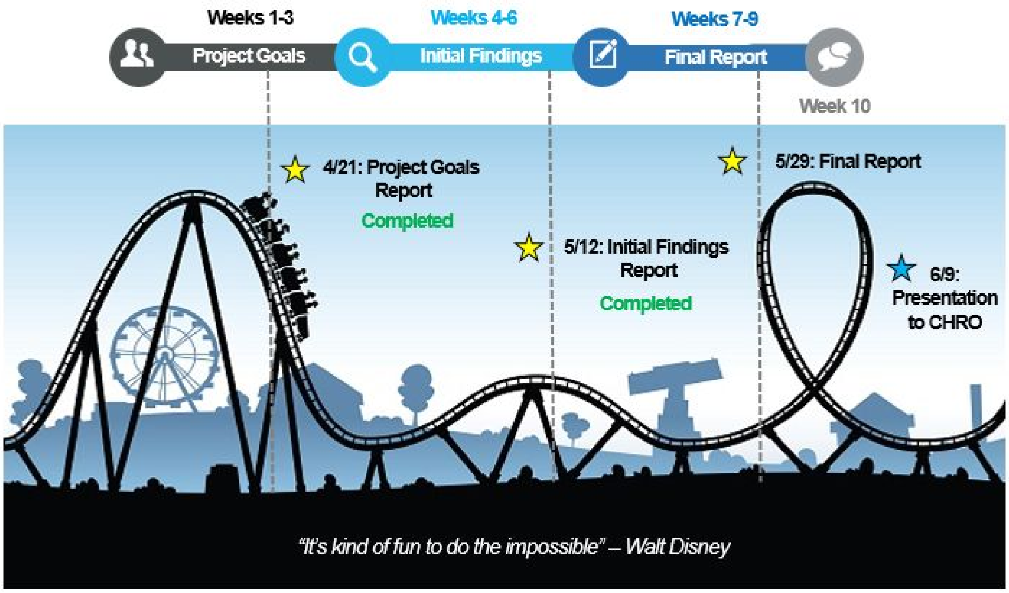
Observations & Learnings	Why it Makes Intuitive Sense
Employees who have lower total working years are more likely to leave the company.	Employees that have less work experience are usually younger and are eager for advancement and progression.

Employees who logged overtime over the course of 1 year were **7 times** more likely to leave the company compared to those who did not.	Employees may be burned out, overworked, or suffer from fatigue.
Employees who are single are **4 times** more likely to leave the company compared to those who are married (1.6 times).	Single employees are more mobile. They don't have a family to support and/or a wife/husband influencing their decisions to stay or leave a company.
Employees who live further away from work are more likely to leave the company.	Cost of gas, inconvenience, and less time with family. Employees may opt for a company that is closer to their home.
Employees that have worked at a number of companies in the past are more likely to leave the company.	One word, "job hoppers"!!!
Employees who are younger are more likely to leave the company.	Employees who are younger are usually more eager for career advancement and progression.
Employees who have low environment satisfaction, job satisfaction, job involvement, relationship satisfaction, or work-life balance survey scores are more likely to leave the company.	Unhappy employees usually leave companies.
Employees who received a low amount of training events in the prior year are more likely to leave the company.	Employees want to be developed and feel appreciated or else employees usually leave.
Employees who have lower daily rates ($) are more likely to leave the company.	Employees usually seek monetary gains and pay raises.
As the tenure of an employee increases, the more likely he/she is to stay with the company.	Higher tenure usually indicates good culture fit, employees that are happy enough to stay, and loyalty to the company.

Project Status

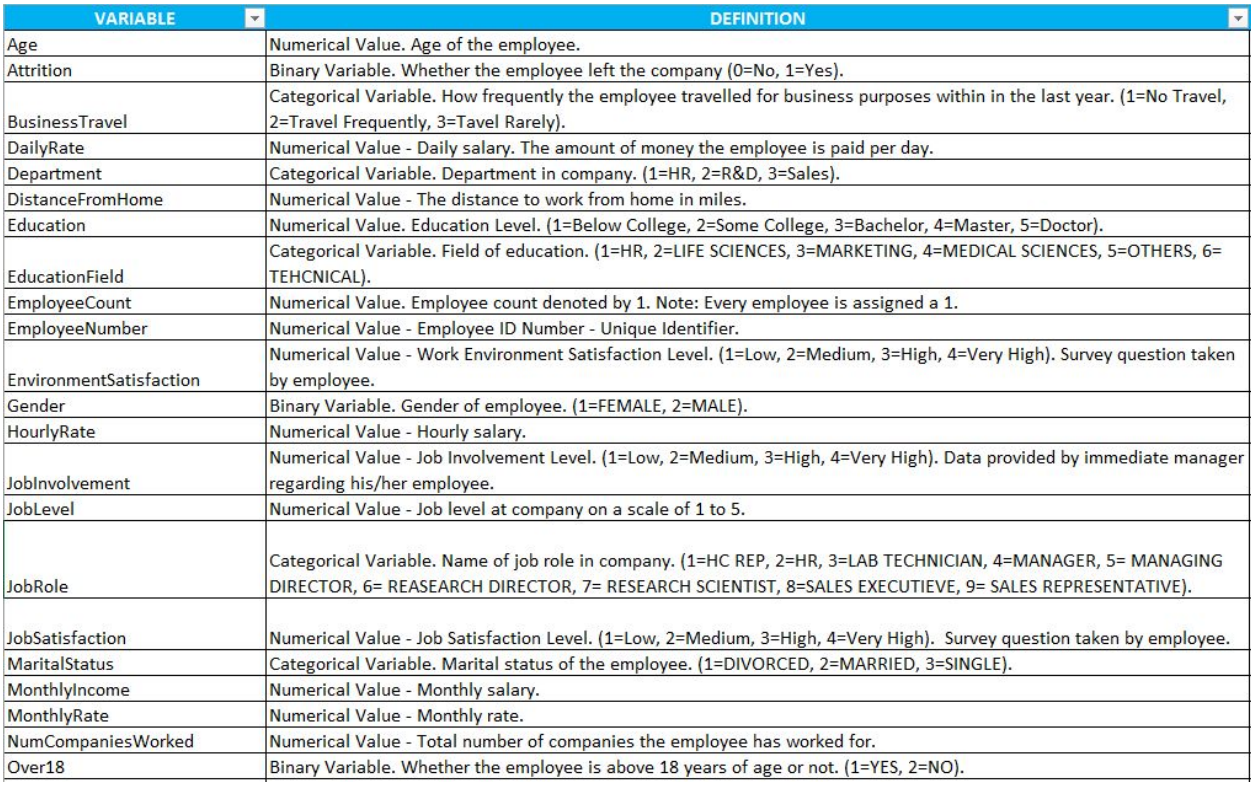
Our project has been following an ambitious 10-week project plan and timeline that we've broken up into 3 phases (project goals, initial findings, and final report). Our team has been meeting at least once or twice a week for 2 -3 hours and communication/collaboration has been excellent via WhatsApp, Google docs, Google Drive, email, and CoLab. The updated timeline is shown below with key milestones identified with stars. We've completed phase 1 and 2 and are now heading into the home stretch. Our primary goal in the next 3 weeks is to create a dashboard/mobile app that can be shared with Corporate HR and business leaders. We also plan on continuing to improve our models using resampling techniques such as SMOTE, experiment with feature engineering, flush out the report, conduct additional analyses (time permitting), and create our final presentation. Recommendations to our CHRO will also be made on how the business can implement the model using a Stay Survey approach. Additionally, programs, policy changes, and initiatives based on our model insights will also be recommended.

As a result, we are "**on track**" to completing our initial goals that were laid out at the beginning of this project. However, if something changes in the next 3 weeks, we will let the CHRO know in our status reports. It's been an honor to work with the Disney team and contribute to the Disney of the future. We look forward to sharing our final results in the upcoming weeks.

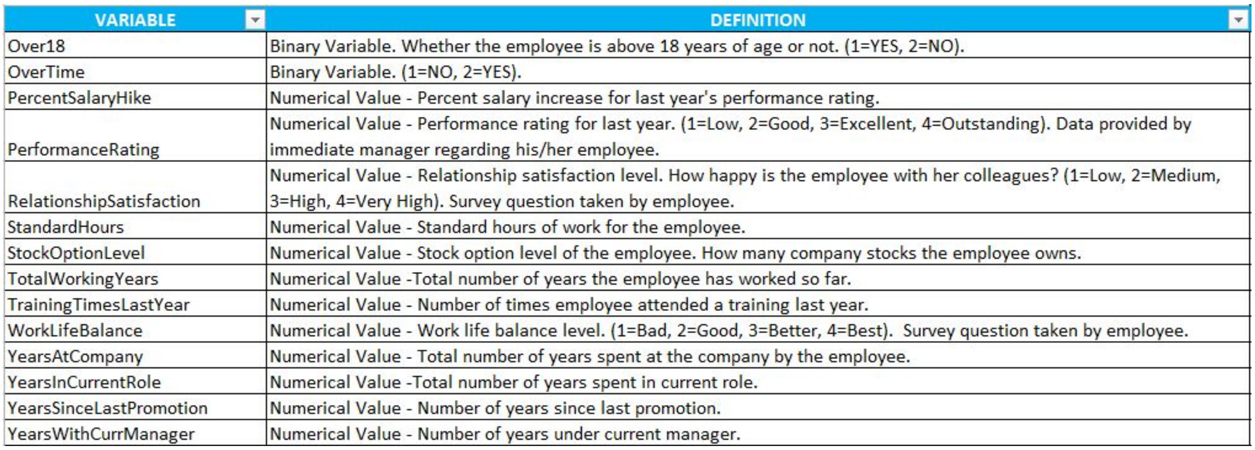

Weeks 1-3
Project Goals
Weeks 4-6
Initial Findings
Weeks 7-9
Final Report
Week 10
4/21: Project Goals Report
Completed
5/12: Initial Findings Report
Completed
5/29: Final Report
6/9: Presentation to CHRO
"It's kind of fun to do the impossible" – Walt Disney

Appendix

Appendix 1: Data Dictionary



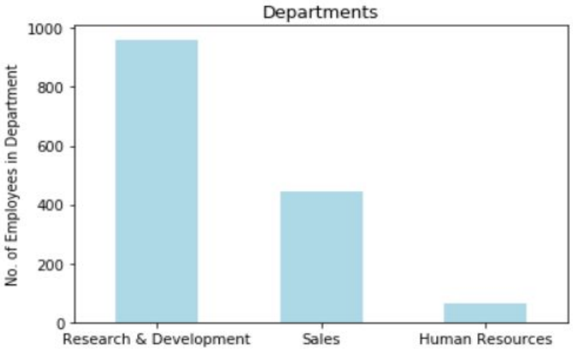
VARIABLE	DEFINITION
Age	Numerical Value. Age of the employee.
Attrition	Binary Variable. Whether the employee left the company (0=No, 1=Yes).
BusinessTravel	Categorical Variable. How frequently the employee travelled for business purposes within in the last year. (1=No Travel, 2=Travel Frequently, 3=Tavel Rarely).
DailyRate	Numerical Value - Daily salary. The amount of money the employee is paid per day.
Department	Categorical Variable. Department in company. (1=HR, 2=R&D, 3=Sales).
DistanceFromHome	Numerical Value - The distance to work from home in miles.
Education	Numerical Value. Education Level. (1=Below College, 2=Some College, 3=Bachelor, 4=Master, 5=Doctor).
EducationField	Categorical Variable. Field of education. (1=HR, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICAL SCIENCES, 5=OTHERS, 6= TEHCNICAL).
EmployeeCount	Numerical Value. Employee count denoted by 1. Note: Every employee is assigned a 1.
EmployeeNumber	Numerical Value - Employee ID Number - Unique Identifier.
EnvironmentSatisfaction	Numerical Value - Work Environment Satisfaction Level. (1=Low, 2=Medium, 3=High, 4=Very High). Survey question taken by employee.
Gender	Binary Variable. Gender of employee. (1=FEMALE, 2=MALE).
HourlyRate	Numerical Value - Hourly salary.
JobInvolvement	Numerical Value - Job Involvement Level. (1=Low, 2=Medium, 3=High, 4=Very High). Data provided by immediate manager regarding his/her employee.
JobLevel	Numerical Value - Job level at company on a scale of 1 to 5.
JobRole	Categorical Variable. Name of job role in company. (1=HC REP, 2=HR, 3=LAB TECHNICIAN, 4=MANAGER, 5= MANAGING DIRECTOR, 6= REASEARCH DIRECTOR, 7= RESEARCH SCIENTIST, 8=SALES EXECUTIEVE, 9= SALES REPRESENTATIVE).
JobSatisfaction	Numerical Value - Job Satisfaction Level. (1=Low, 2=Medium, 3=High, 4=Very High). Survey question taken by employee.
MaritalStatus	Categorical Variable. Marital status of the employee. (1=DIVORCED, 2=MARRIED, 3=SINGLE).
MonthlyIncome	Numerical Value - Monthly salary.
MonthlyRate	Numerical Value - Monthly rate.
NumCompaniesWorked	Numerical Value - Total number of companies the employee has worked for.
Over18	Binary Variable. Whether the employee is above 18 years of age or not. (1=YES, 2=NO).

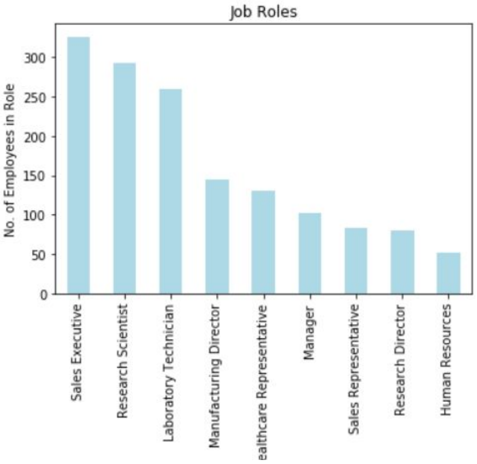


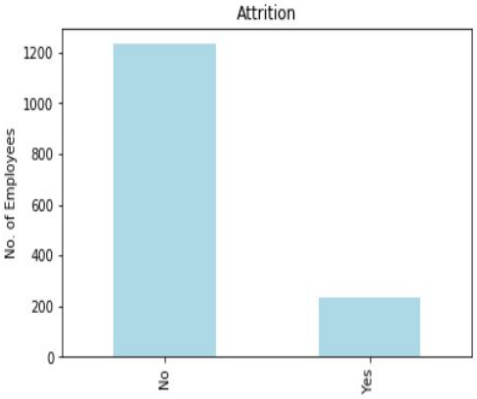
VARIABLE	DEFINITION
Over18	Binary Variable. Whether the employee is above 18 years of age or not. (1=YES, 2=NO).
OverTime	Binary Variable. (1=NO, 2=YES).
PercentSalaryHike	Numerical Value - Percent salary increase for last year's performance rating.
PerformanceRating	Numerical Value - Performance rating for last year. (1=Low, 2=Good, 3=Excellent, 4=Outstanding). Data provided by immediate manager regarding his/her employee.
RelationshipSatisfaction	Numerical Value - Relationship satisfaction level. How happy is the employee with her colleagues? (1=Low, 2=Medium, 3=High, 4=Very High). Survey question taken by employee.
StandardHours	Numerical Value - Standard hours of work for the employee.
StockOptionLevel	Numerical Value - Stock option level of the employee. How many company stocks the employee owns.
TotalWorkingYears	Numerical Value -Total number of years the employee has worked so far.
TrainingTimesLastYear	Numerical Value - Number of times employee attended a training last year.
WorkLifeBalance	Numerical Value - Work life balance level. (1=Bad, 2=Good, 3=Better, 4=Best). Survey question taken by employee.
YearsAtCompany	Numerical Value - Total number of years spent at the company by the employee.
YearsInCurrentRole	Numerical Value -Total number of years spent in current role.
YearsSinceLastPromotion	Numerical Value - Number of years since last promotion.
YearsWithCurrManager	Numerical Value - Number of years under current manager.

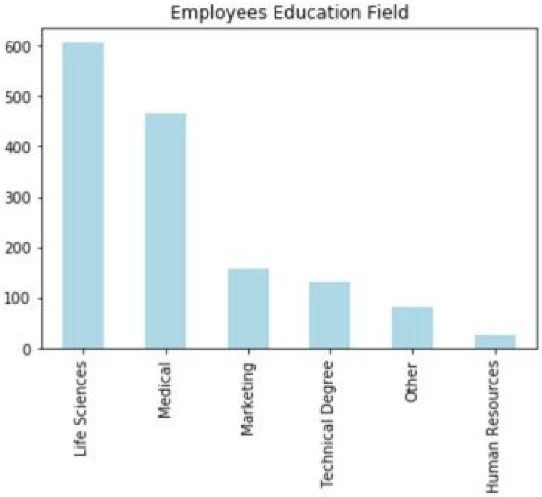
Appendix 2: Variable Distributions of Categorical & Numeric Variables

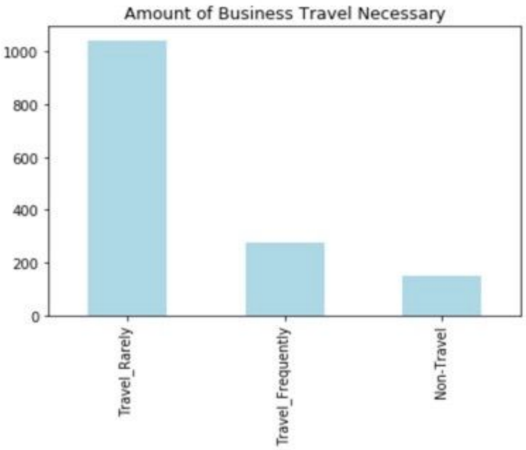
Bar plots and histograms were generated below to get a better understanding of the distributions of each variable.

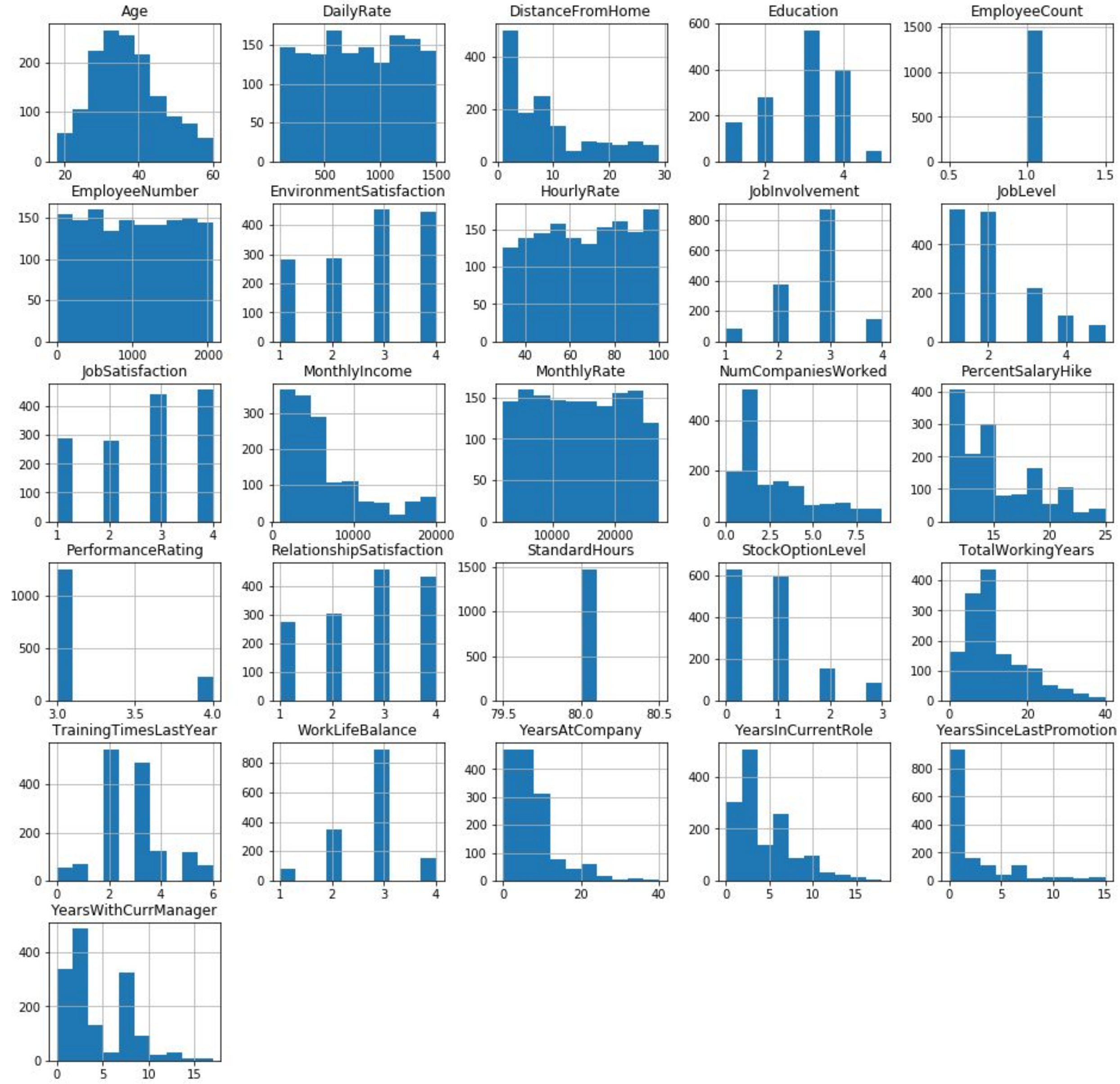






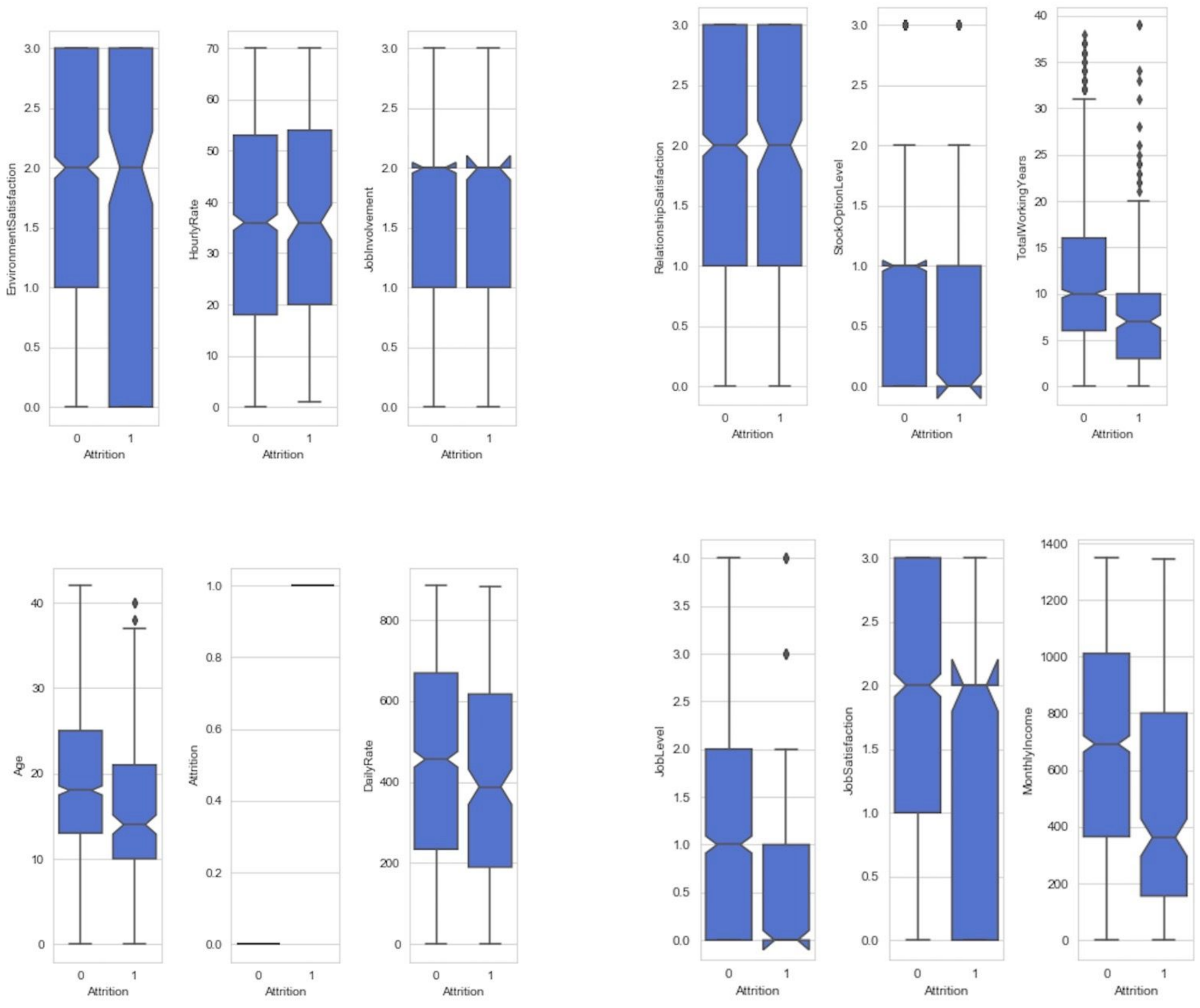


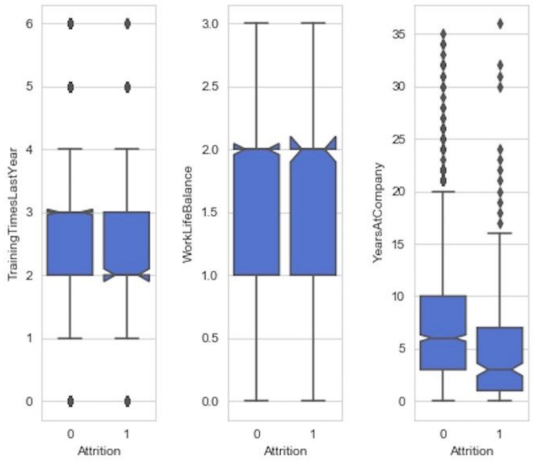


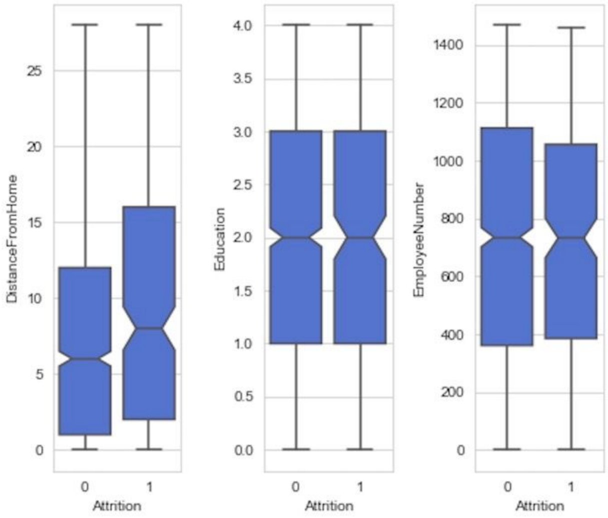

Age
0
100
200
20
40
60
DailyRate
0
50
100
150
500
1000
1500
DistanceFromHome
0
200
400
0
10
20
30
Education
0
200
400
600
2
4
EmployeeCount
0
500
1000
1500
0.5
1.0
1.5
EmployeeNumber
0
50
100
150
0
1000
2000
EnvironmentSatisfaction
0
100
200
300
400
1
2
3
4
HourlyRate
0
50
100
150
40
60
80
100
JobInvolvement
0
200
400
600
800
1
2
3
4
JobLevel
0
200
400
2
4
JobSatisfaction
0
100
200
300
400
1
2
3
4
MonthlyIncome
0
100
200
300
0
10000
20000
MonthlyRate
0
50
100
150
10000
20000
NumCompaniesWorked
0
200
400
0.0
2.5
5.0
7.5
PercentSalaryHike
0
100
200
300
400
15
20
25
PerformanceRating
0
500
1000
3.0
3.5
4.0
RelationshipSatisfaction
0
100
200
300
400
1
2
3
4
StandardHours
0
500
1000
1500
79.5
80.0
80.5
StockOptionLevel
0
200
400
600
0
1
2
3
TotalWorkingYears
0
100
200
300
400
0
20
40
TrainingTimesLastYear
0
200
400
0
2
4
6
WorkLifeBalance
0
200
400
600
800
1
2
3
4
YearsAtCompany
0
100
200
300
400
0
20
40
YearsInCurrentRole
0
200
400
0
5
10
15
YearsSinceLastPromotion
0
200
400
600
800
0
5
10
15
YearsWithCurrManager
0
200
400
0
5
10
15

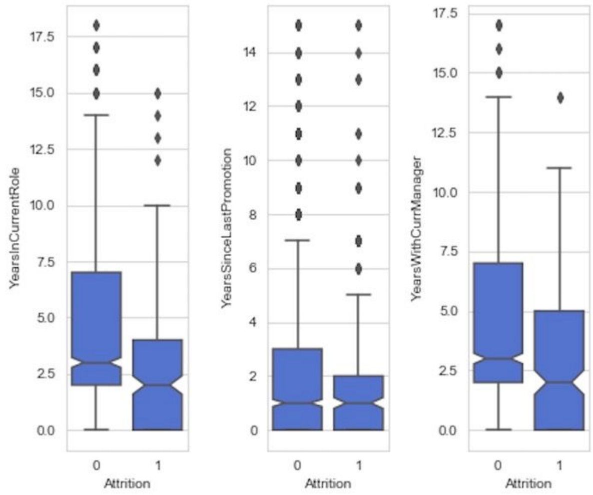
Appendix 3: Notched box plots of numeric variables vs. attrition.

*The notch displays a confidence interval around the median which is normally based on the median +/- 1.58*IQR/sqrt(n), which allows us to visually compare if the medians differ. For instance, if the notch does not overlap, it means that the median differs and indicates that the variable is potentially a strong predictor to include in our models*.



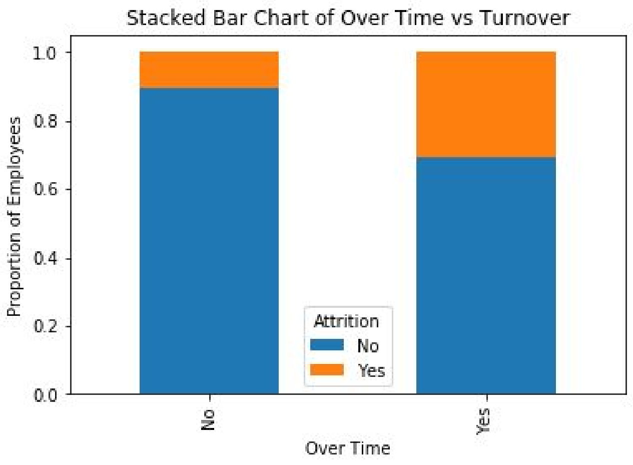

TrainingTimesLastYear
0
1
2
3
4
5
6
Attrition
WorkLifeBalance
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Attrition
YearsAtCompany
0
5
10
15
20
25
30
35
Attrition

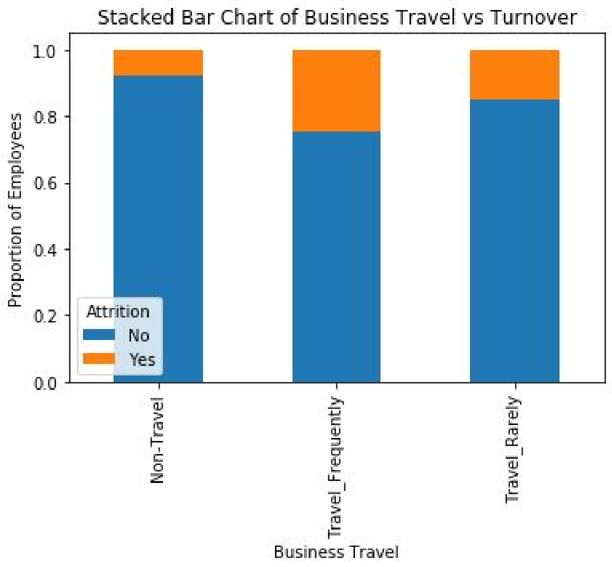

DistanceFromHome
0
5
10
15
20
25
Attrition
Education
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Attrition
EmployeeNumber
0
200
400
600
800
1000
1200
1400
Attrition

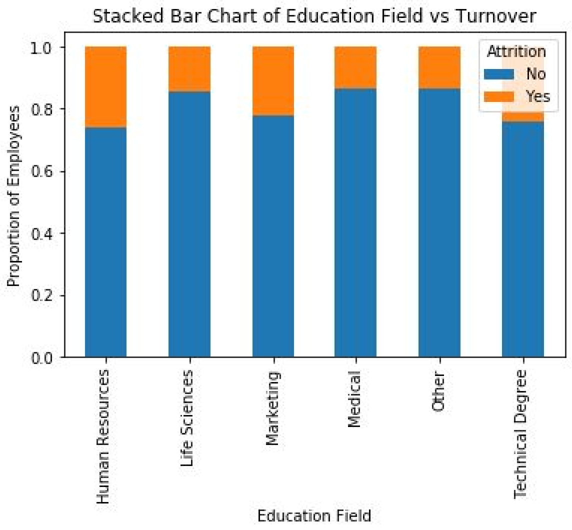

YearsInCurrentRole
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Attrition
YearsSinceLastPromotion
0
2
4
6
8
10
12
14
Attrition
YearsWithCurrManager
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Attrition

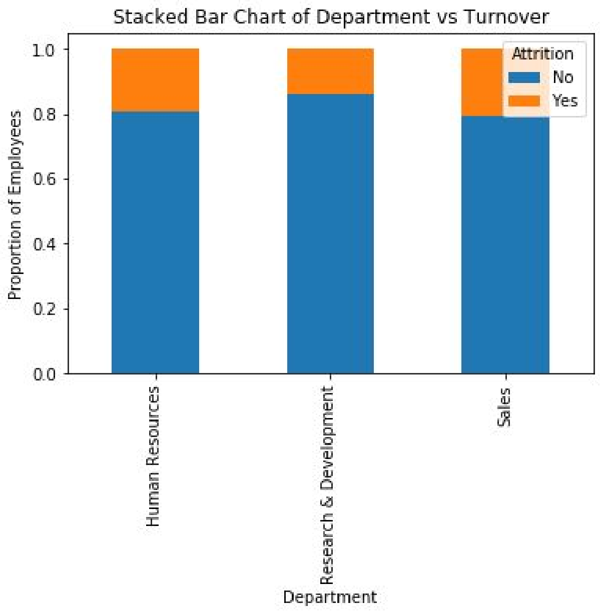
Appendix 4: Stacked bar plots of categorical variables vs. attrition.

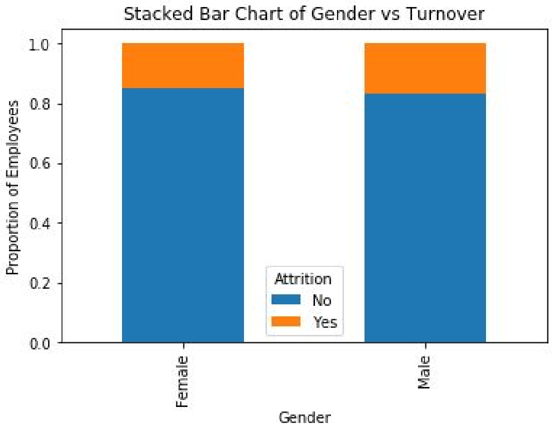
Stacked bar plots were produced below to see if there was variability between whether a person stayed or left the company. If there was high variability, this indicates that the variable is potentially a strong predictor to include in our models.

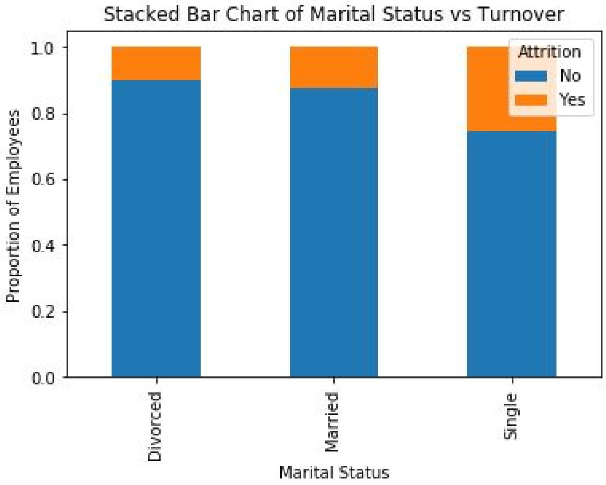






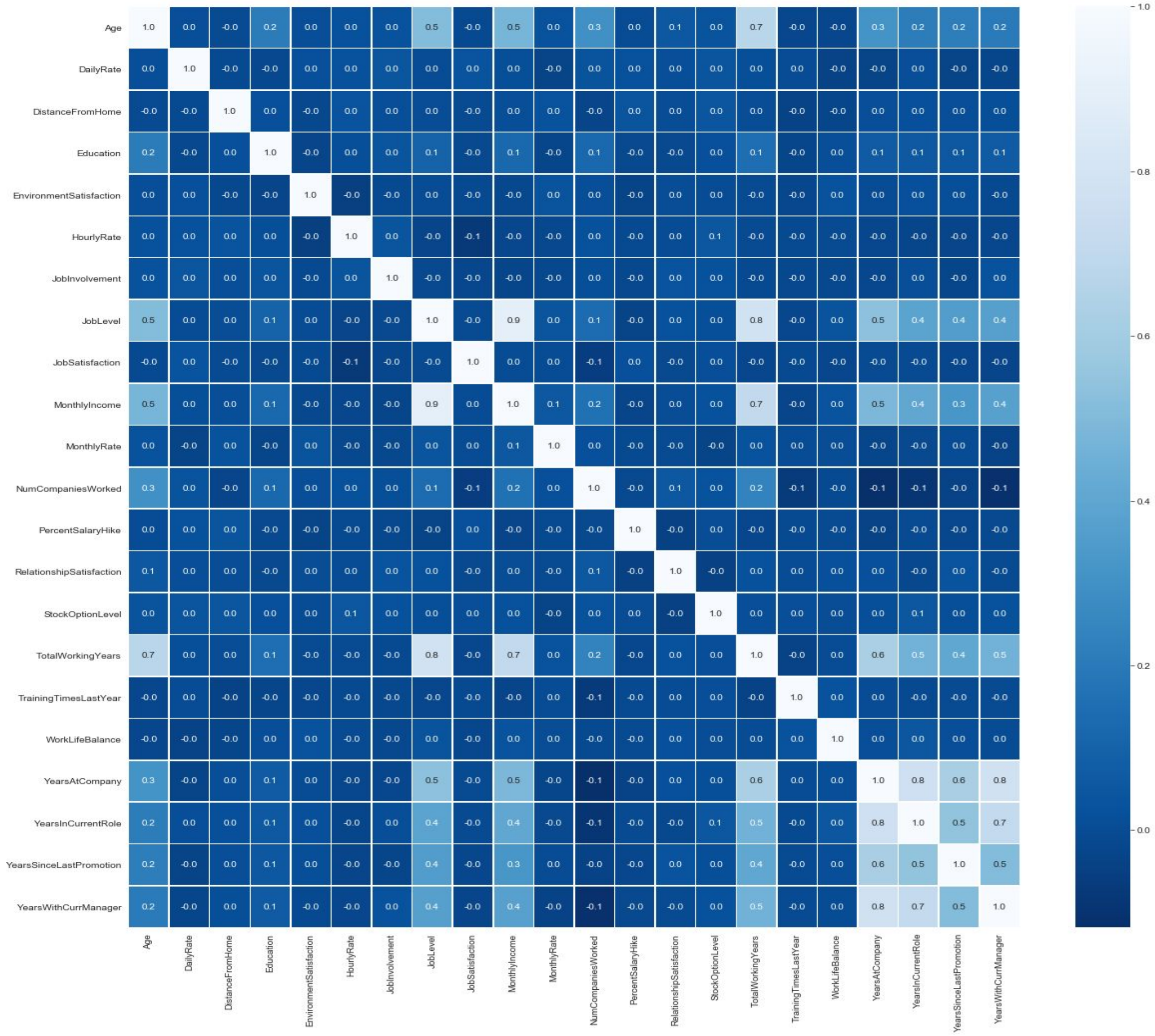



Stacked Bar Chart of Gender vs Turnover
Proportion of Employees
1.0
0.8
0.6
0.4
0.2
0.0
Attrition
No
Yes
Female
Male
Gender


Stacked Bar Chart of Marital Status vs Turnover
Proportion of Employees
1.0
0.8
0.6
0.4
0.2
0.0
Attrition
No
Yes
Divorced
Married
Single
Marital Status

Appendix 5: Correlation Matrix

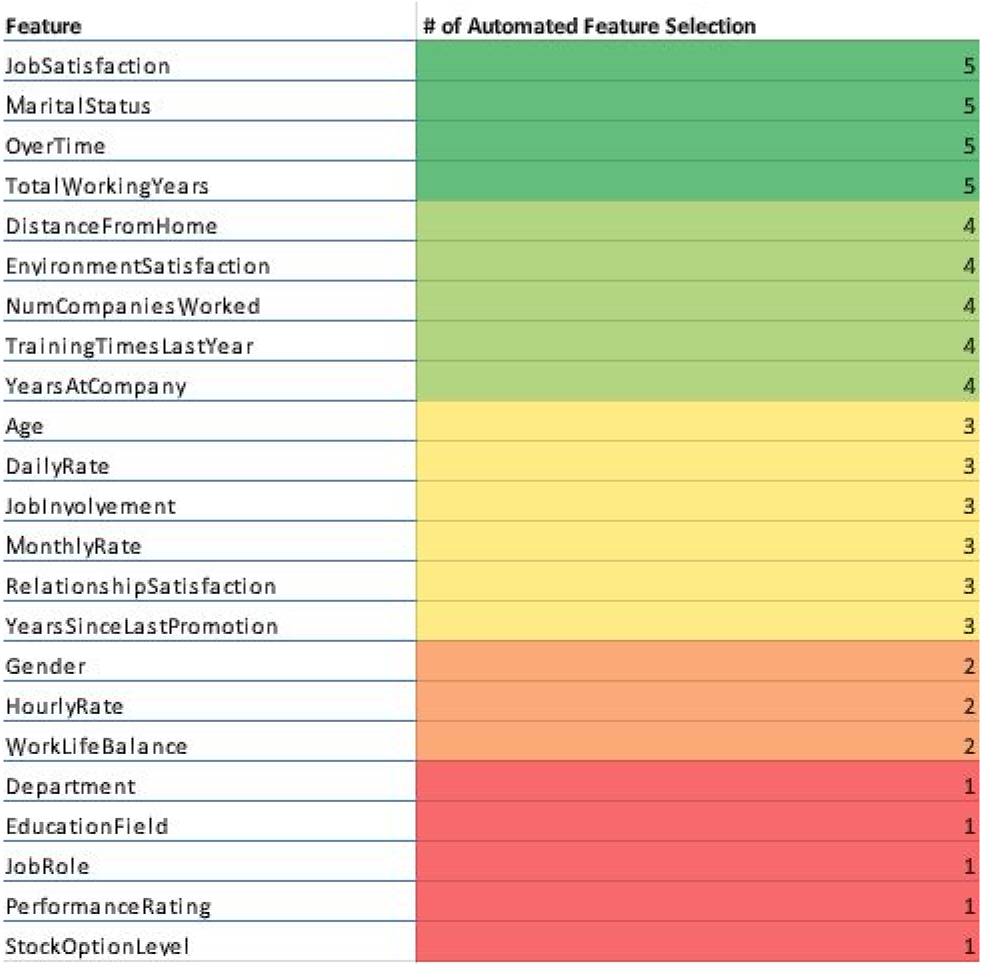
A correlation matrix helped us determine if some of the predictors are highly correlated with each other so that only one of those variables are included in the automated feature selection techniques. Using this approach also removed a variable from each pair of collinear variables so that automated feature selection techniques didn't end up choosing similar and collinear variables. Removing similar variables also helped eliminate multicollinearity concerns. The correlation matrix revealed several variables that had correlations greater than 0.70



	Age	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfaction	MonthlyIncome	MonthlyRate	NumCompaniesWorked	PercentSalaryHike	RelationshipSatisfaction	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
Age	1.0	0.0	-0.0	0.2	0.0	0.0	0.0	0.5	-0.0	0.5	0.0	0.3	0.0	0.1	0.0	0.7	-0.0	-0.0	0.3	0.2	0.2	0.2
DailyRate	0.0	1.0	-0.0	-0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	0.0	0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	-0.0	-0.0
DistanceFromHome	-0.0	-0.0	1.0	0.0	-0.0	0.0	0.0	0.0	-0.0	0.0	0.0	-0.0	0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	0.0	0.0
Education	0.2	-0.0	0.0	1.0	-0.0	0.0	0.0	0.1	-0.0	0.1	-0.0	0.1	-0.0	-0.0	0.0	0.1	-0.0	0.0	0.1	0.1	0.1	0.1
EnvironmentSatisfaction	0.0	0.0	-0.0	-0.0	1.0	-0.0	-0.0	0.0	-0.0	-0.0	0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	-0.0
HourlyRate	0.0	0.0	0.0	0.0	-0.0	1.0	0.0	-0.0	-0.1	-0.0	-0.0	0.0	-0.0	0.0	0.1	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0
JobInvolvement	0.0	0.0	0.0	0.0	-0.0	0.0	1.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	0.0
JobLevel	0.5	0.0	0.0	0.1	0.0	-0.0	-0.0	1.0	-0.0	0.9	0.0	0.1	-0.0	0.0	0.0	0.8	-0.0	0.0	0.5	0.4	0.4	0.4
JobSatisfaction	-0.0	0.0	-0.0	-0.0	-0.0	-0.1	-0.0	-0.0	1.0	0.0	0.0	-0.1	0.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0
MonthlyIncome	0.5	0.0	0.0	0.1	-0.0	-0.0	-0.0	0.9	0.0	1.0	0.1	0.2	-0.0	0.0	0.0	0.7	-0.0	0.0	0.5	0.4	0.3	0.4
MonthlyRate	0.0	-0.0	0.0	-0.0	0.0	-0.0	-0.0	0.0	0.0	0.1	1.0	0.0	-0.0	-0.0	-0.0	0.0	0.0	0.0	-0.0	-0.0	0.0	-0.0
NumCompaniesWorked	0.3	0.0	-0.0	0.1	0.0	0.0	0.0	0.1	-0.1	0.2	0.0	1.0	-0.0	0.1	0.0	0.2	-0.1	-0.0	-0.1	-0.1	-0.0	-0.1
PercentSalaryHike	0.0	0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.0	-0.0	-0.0	1.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0
RelationshipSatisfaction	0.1	0.0	0.0	-0.0	0.0	0.0	0.0	0.0	-0.0	0.0	-0.0	0.1	-0.0	1.0	-0.0	0.0	0.0	0.0	0.0	-0.0	0.0	-0.0
StockOptionLevel	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	-0.0	0.0	0.0	-0.0	1.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
TotalWorkingYears	0.7	0.0	0.0	0.1	-0.0	-0.0	-0.0	0.8	-0.0	0.7	0.0	0.2	-0.0	0.0	0.0	1.0	-0.0	0.0	0.6	0.5	0.4	0.5
TrainingTimesLastYear	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	0.0	-0.1	-0.0	0.0	0.0	-0.0	1.0	0.0	0.0	-0.0	-0.0	-0.0
WorkLifeBalance	-0.0	-0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0	-0.0	0.0	0.0	-0.0	-0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
YearsAtCompany	0.3	-0.0	0.0	0.1	0.0	-0.0	-0.0	0.5	-0.0	0.5	-0.0	-0.1	-0.0	0.0	0.0	0.6	0.0	0.0	1.0	0.8	0.6	0.8
YearsInCurrentRole	0.2	0.0	0.0	0.1	0.0	-0.0	0.0	0.4	-0.0	0.4	-0.0	-0.1	-0.0	-0.0	0.1	0.5	-0.0	0.0	0.8	1.0	0.5	0.7
YearsSinceLastPromotion	0.2	-0.0	0.0	0.1	0.0	-0.0	-0.0	0.4	-0.0	0.3	0.0	-0.0	-0.0	0.0	0.0	0.4	-0.0	0.0	0.6	0.5	1.0	0.5
YearsWithCurrManager	0.2	-0.0	0.0	0.1	-0.0	-0.0	0.0	0.4	-0.0	0.4	-0.0	-0.1	-0.0	-0.0	0.0	0.5	-0.0	0.0	0.8	0.7	0.5	1.0

Appendix 6: Summary of Automated Variable Selection Techniques

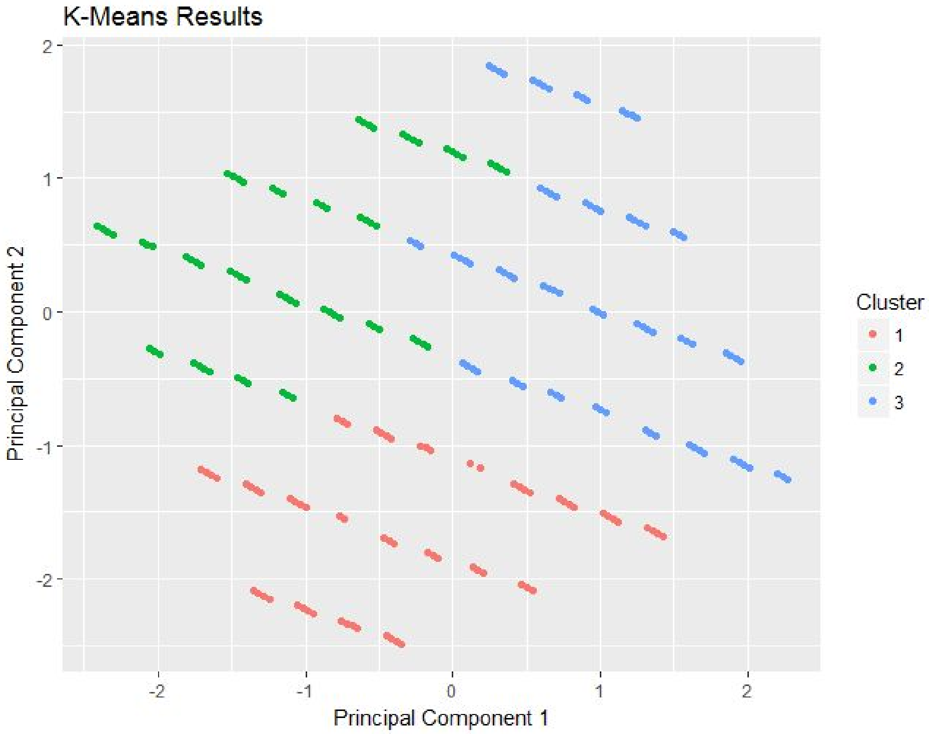
The People Insights team utilized 5 automated variable selection techniques to help select which features were important (e.g., univariate feature selection, recursive feature elimination , tree-based feature selection, a decision tree, and lasso embedded method on the training data set) . The table below shows the amount of times each variable appeared using the techniques above. Variables that appeared most frequently were incorporated into the model.



Feature	# of Automated Feature Selection
JobSatisfaction	5
MaritalStatus	5
OverTime	5
TotalWorkingYears	5
DistanceFromHome	4
EnvironmentSatisfaction	4
NumCompaniesWorked	4
TrainingTimesLastYear	4
YearsAtCompany	4
Age	3
DailyRate	3
JobInvolvement	3
MonthlyRate	3
RelationshipSatisfaction	3
YearsSinceLastPromotion	3
Gender	2
HourlyRate	2
WorkLifeBalance	2
Department	1
EducationField	1
JobRole	1
PerformanceRating	1
StockOptionLevel	1

Appendix 7: Cluster Analysis

K-means clustering is a nonhierarchical or partition based method that splits the data into “predetermined” number clusters so that items within each cluster are similar to each other and items from other clusters are dissimilar (Izenman, 2008). The first 2 principal components explain 0.5927590 of the variation in the data. The following visual below shows the cluster results from k-means. As you can see, there are 3 clusters: job hoppers, disengaged low performers, and engaged high performers.

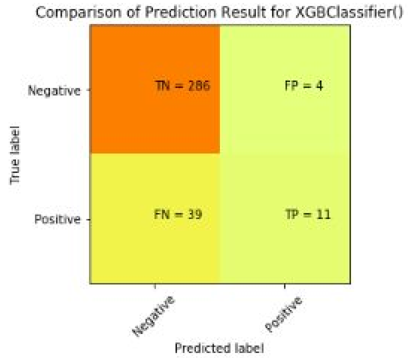


1:Job Hoppers 2:Disengaged Low Performers 3:Engaged High Performers

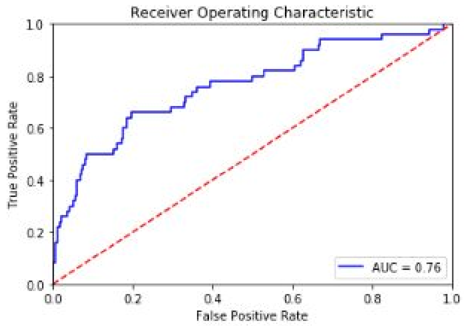
Appendix 8: Model Results

XGBoost

Boosting is an ensemble learning method that provides another approach for improving the predictions resulting from a decision tree by fitting each tree on an altered version of the original dataset (James, et al., 2013). In other words, trees are grown sequentially (e.g., each tree is grown using information from previously grown trees). XGBoost is similar to GBM, in the sense that it follows the same principle of gradient boosting. However, XGBoost uses different modeling parameters, incorporates parallel processing, and uses a more regularized model formalization to control over-fitting, which often results in better performance.



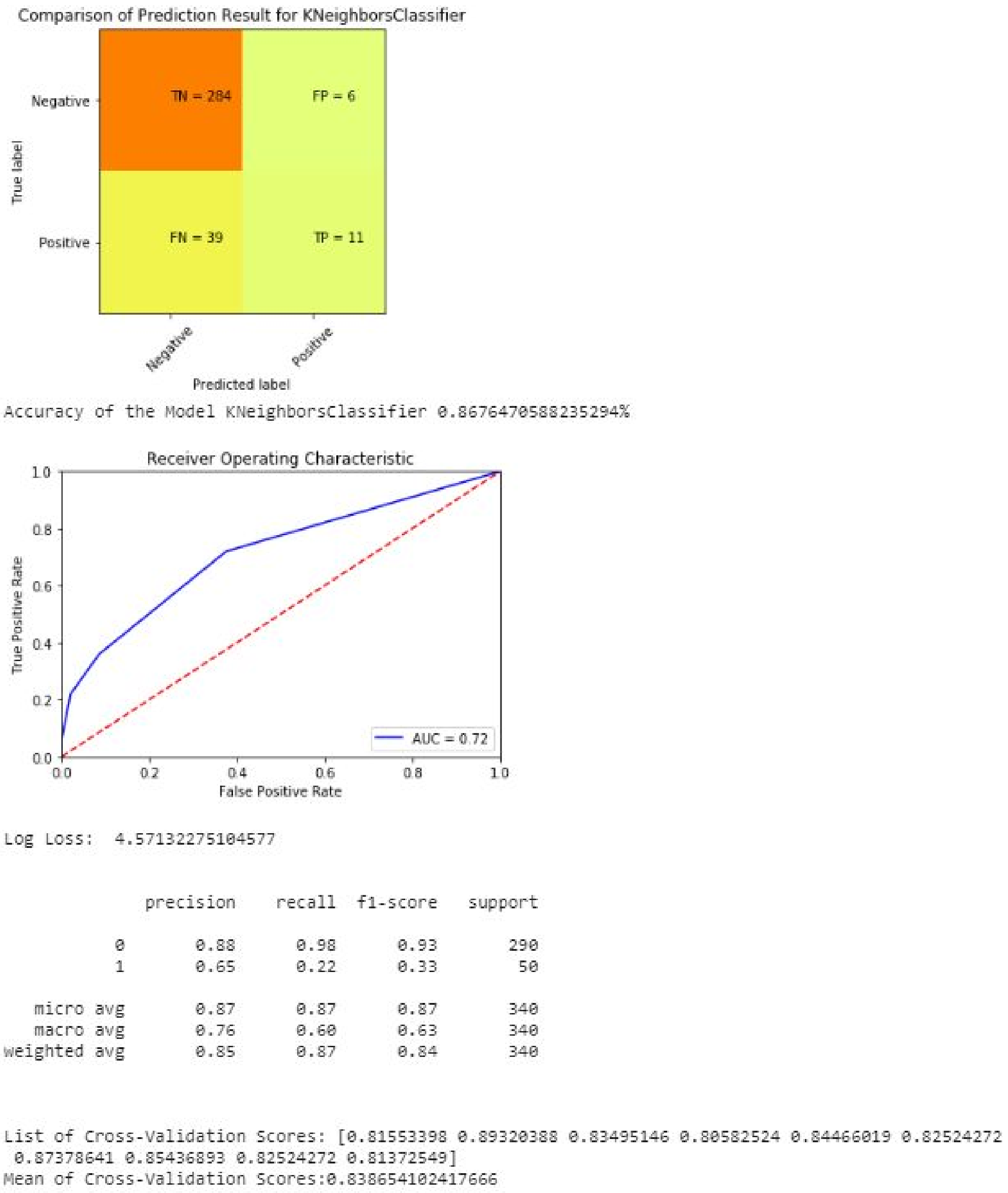
```
Accuracy of the Model XGBClassifier() 0.8735294117647059%
```



```
Log Loss:  4.368148774620236

             precision    recall  f1-score   support

          0       0.88      0.99      0.93       290
          1       0.73      0.22      0.34        50

  micro avg       0.87      0.87      0.87       340
  macro avg       0.81      0.60      0.63       340
weighted avg      0.86      0.87      0.84       340


List of Cross-Validation Scores: [0.85436893 0.89320388 0.82524272 0.80582524 0.88349515 0.82524272
 0.86407767 0.86407767 0.84466019 0.84313725]
Mean of Cross-Validation Scores:0.8503331429659242
```

K Nearest Neighbors Model 1

KNN is a non-parametric method that applies Bayes rule by classifying a given observation to the class with the highest estimated probability based on a similarity measure (e.g., distance functions) (James, et al., 2013).



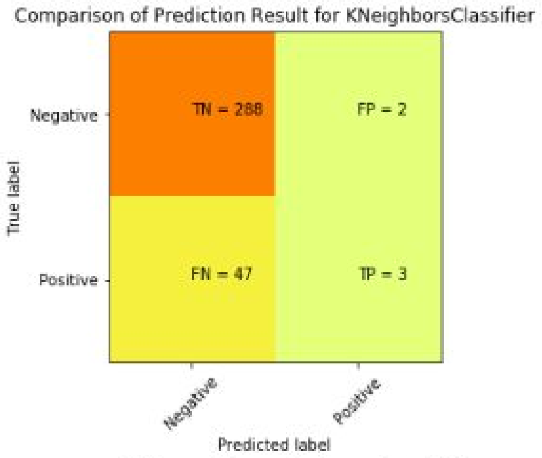
```
Accuracy of the Model KNeighborsClassifier 0.8676470588235294%
```

```
Log Loss:  4.57132275104577
```

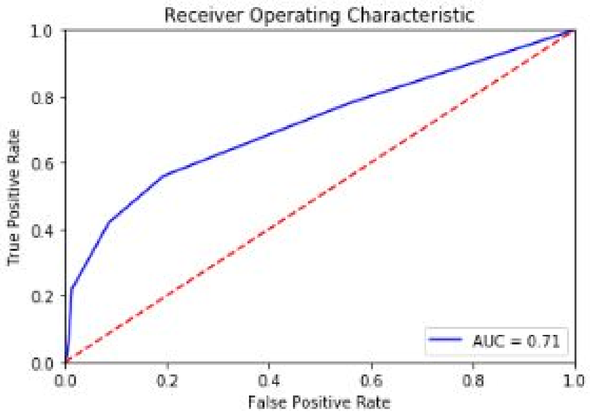
```
             precision    recall  f1-score   support

          0       0.88      0.98      0.93       290
          1       0.65      0.22      0.33        50

  micro avg       0.87      0.87      0.87       340
  macro avg       0.76      0.60      0.63       340
weighted avg      0.85      0.87      0.84       340
```

```
List of Cross-Validation Scores: [0.81553398 0.89320388 0.83495146 0.80582524 0.84466019 0.82524272
 0.87378641 0.85436893 0.82524272 0.81372549]
Mean of Cross-Validation Scores:0.838654102417666
```

K Nearest Neighbors Model 2



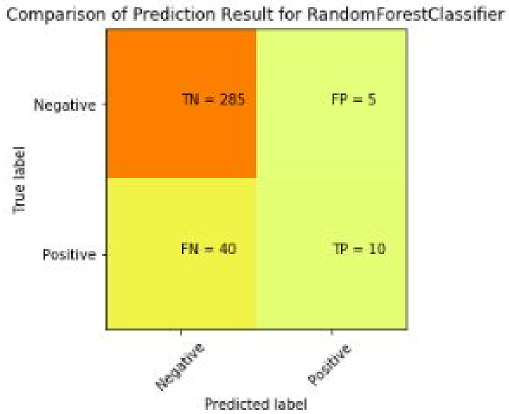
```
Accuracy of the Model KNeighborsClassifier 0.8558823529411764%
```



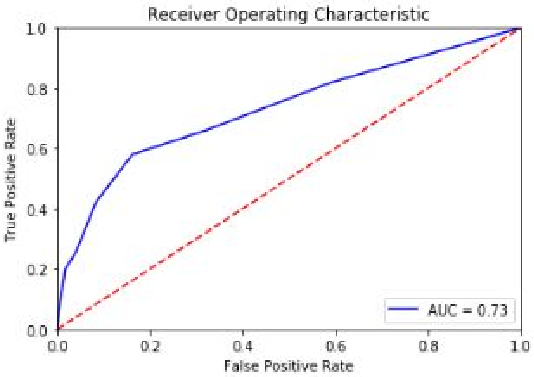
```
Log Loss:  4.977651889839659

             precision    recall  f1-score   support

           0       0.86      0.99      0.92       290
           1       0.60      0.06      0.11        50

   micro avg       0.86      0.86      0.86       340
   macro avg       0.73      0.53      0.52       340
weighted avg       0.82      0.86      0.80       340


List of Cross-Validation Scores: [0.84466019 0.87378641 0.85436893 0.84466019 0.84466019 0.85436893
 0.86407767 0.85436893 0.84466019 0.84313725]
Mean of Cross-Validation Scores:0.8522748905387397
```

Random Forest

Random forest is an ensemble learning method that provides an improvement over bagged trees by incorporating a small tweak that decorrelates the trees and then averages them (e.g., forces each split to only consider a subset of predictors and will not consider strong predictors so that other predictors will have more of a chance (James, et al., 2013)).



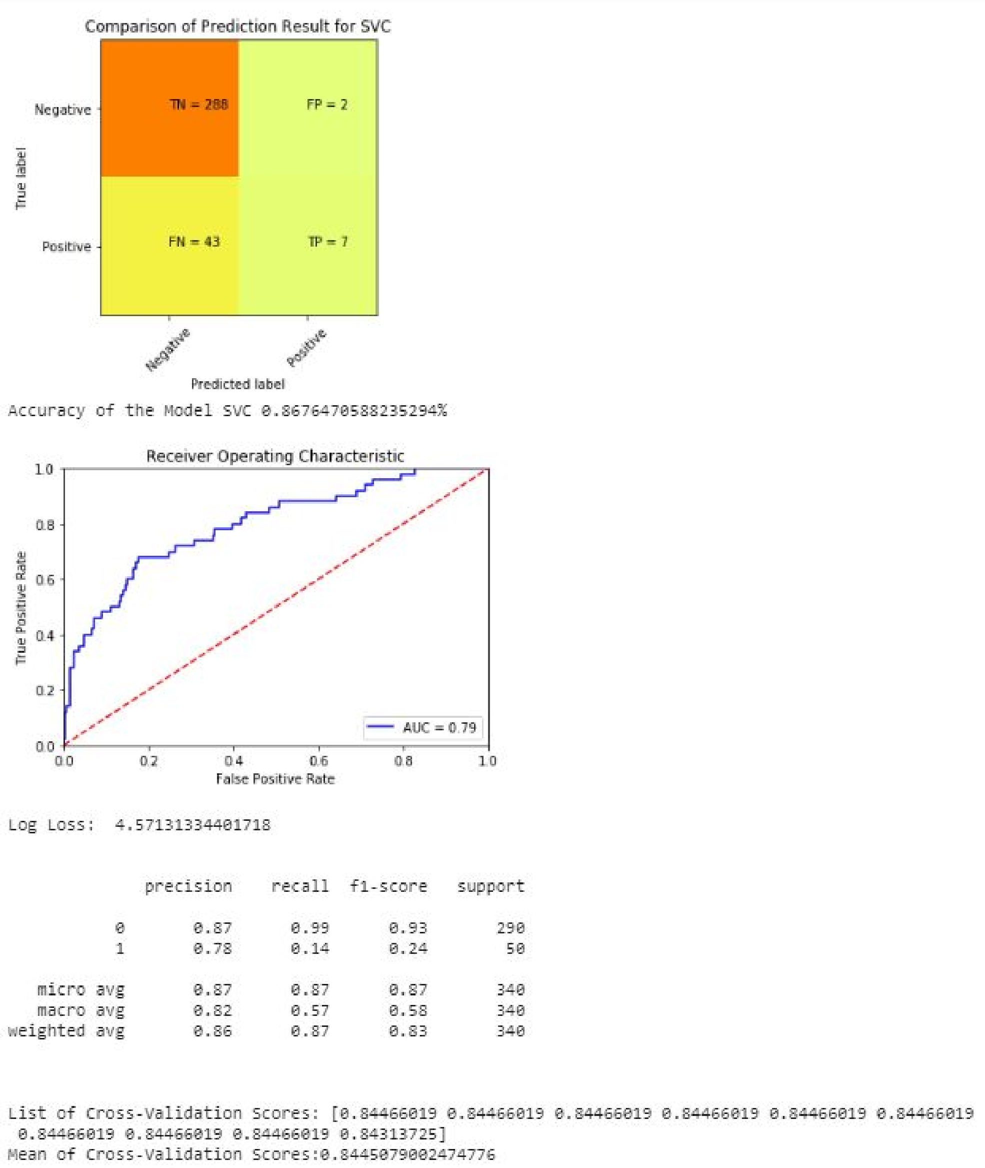
```
Accuracy of the Model RandomForestClassifier 0.8676470588235294%
```



```
Log Loss:  4.5713203992886235

              precision    recall  f1-score   support

           0       0.88      0.98      0.93       290
           1       0.67      0.20      0.31        50

   micro avg       0.87      0.87      0.87       340
   macro avg       0.77      0.59      0.62       340
weighted avg       0.85      0.87      0.84       340


List of Cross-Validation Scores: [0.83495146 0.85436893 0.83495146 0.83495146 0.83495146 0.80582524
 0.88349515 0.86407767 0.83495146 0.87254902]
Mean of Cross-Validation Scores:0.8455073291452502
```

Support Vector Machines

Support vector machine (SVM) is often considered one of the best "out of the box" classifiers (James, et al., 2013). A support vector classifier model is a linear classifier that allows some training observations to be misclassified in order to improve classification for the remaining observations (James, et al., 2013).

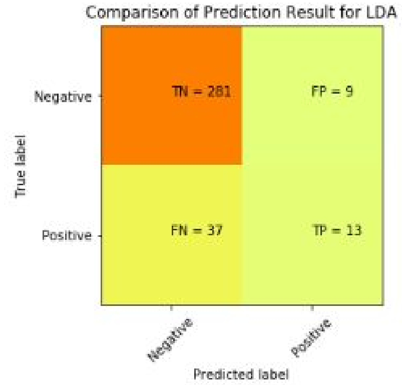


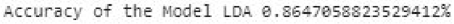
```
Accuracy of the Model SVC 0.8676470588235294%
```

```
Log Loss:  4.57131334401718


              precision    recall  f1-score   support

           0       0.87      0.99      0.93       290
           1       0.78      0.14      0.24        50

   micro avg       0.87      0.87      0.87       340
   macro avg       0.82      0.57      0.58       340
weighted avg       0.86      0.87      0.83       340


List of Cross-Validation Scores: [0.84466019 0.84466019 0.84466019 0.84466019 0.84466019 0.84466019
 0.84466019 0.84466019 0.84466019 0.84313725]
Mean of Cross-Validation Scores:0.8445079002474776
```

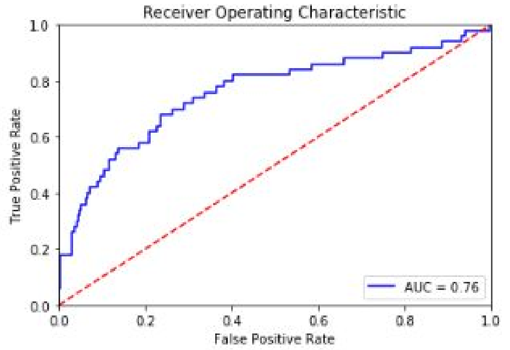
Linear Discriminant Analysis

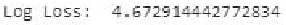
LDA is very similar in form to logistic regression (distributions are assumed to be normal), except it models the distribution of the predictors separately in each of the response classes and then applies Bayes theorem (James, et al., 2013). This model also uses Gaussian densities, which arises when we assume that the classes have a common covariance matrix, and assumes a linear decision boundary (Hastie, et al., 2009).



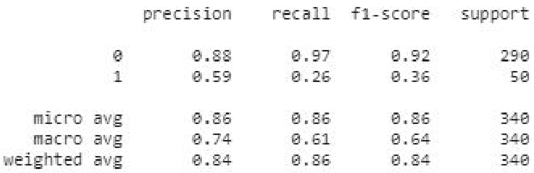


```
Accuracy of the Model LDA 0.8647058823529412%
```

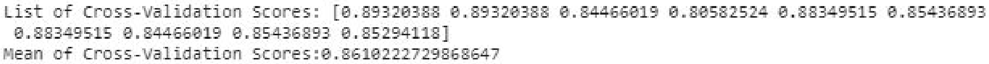




```
Log Loss:  4.672914442772834
```



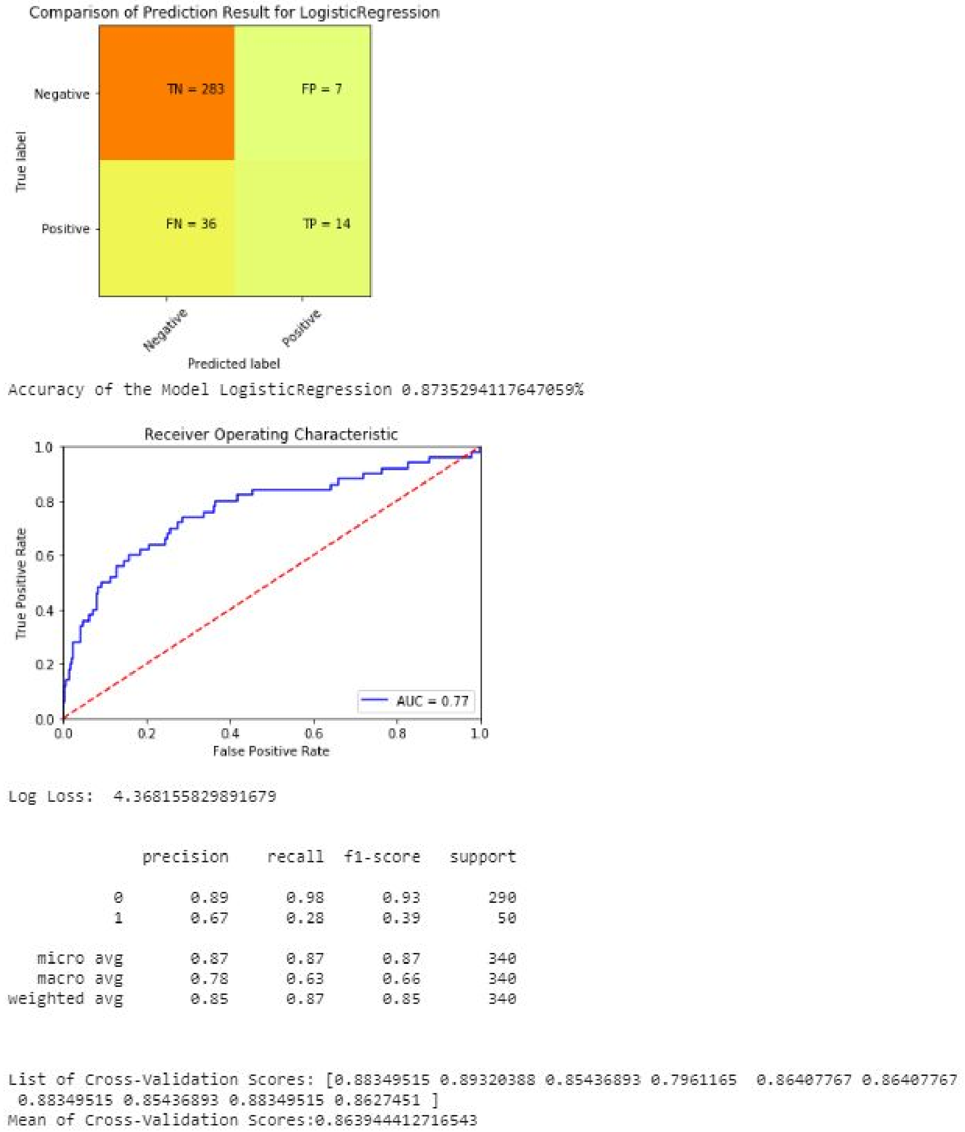
```
             precision    recall  f1-score   support

           0       0.88      0.97      0.92       290
           1       0.59      0.26      0.36        50

   micro avg       0.86      0.86      0.86       340
   macro avg       0.74      0.61      0.64       340
weighted avg       0.84      0.86      0.84       340
```



```
List of Cross-Validation Scores: [0.89320388 0.89320388 0.84466019 0.80582524 0.88349515 0.85436893
 0.88349515 0.84466019 0.85436893 0.85294118]
Mean of Cross-Validation Scores:0.8610222729868647
```

Logistic Regression

Logistic regression models the probability that the response variable belongs to a specific category and assumes a linear decision boundary (James, Witten, Hastie, & Tibshirani, 2013). For instance, it models the probabilities of the K classes using linear functions in x, while also ensuring that they sum to 1 and remain in-between 0 and 1 (Hastie, Tibshirani, & Friedman, 2009). This is accomplished using the logistic function and maximum likelihood, which is used to fit the model (James, et al., 2013).

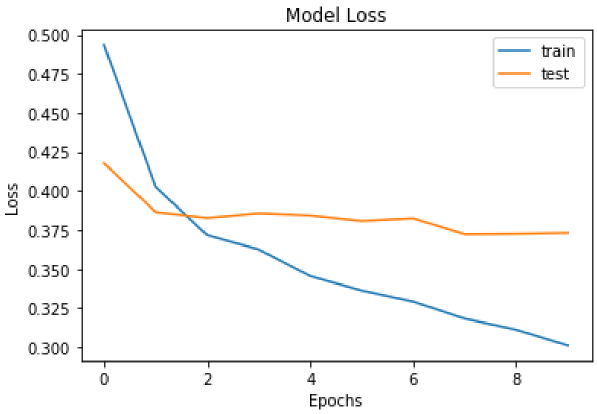


```
Accuracy of the Model LogisticRegression 0.8735294117647059%
```

```
Log Loss:  4.368155829891679

              precision    recall  f1-score   support

           0       0.89      0.98      0.93       290
           1       0.67      0.28      0.39        50

   micro avg       0.87      0.87      0.87       340
   macro avg       0.78      0.63      0.66       340
weighted avg       0.85      0.87      0.85       340


List of Cross-Validation Scores: [0.88349515 0.89320388 0.85436893 0.7961165  0.86407767 0.86407767
 0.88349515 0.85436893 0.88349515 0.8627451 ]
Mean of Cross-Validation Scores:0.863944412716543
```

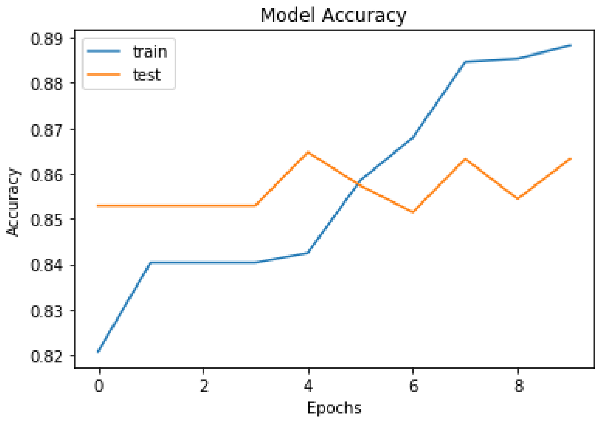
Neural Network

Neural network (aka: single hidden layer back-propagation network) is a nonlinear statistical model that is basically a nonlinear generalization of a linear model (Hastie, et al., 2009). It contains inputs, a hidden layer, and outputs that are typically represented by a network diagram (Hastie, et al., 2009). Additionally, neural network has unknown parameters called weights that introduce nonlinearities where needed (Hastie, et al., 2009).



Average loss and accuracy on train

```
[0.2866899906204816, 0.8947750363709792]
```



Average loss and accuracy on test

```
[0.3730934977531433, 0.8632352934164159]
```

References

1. Izenman, A. J. (2008). Modern multivariate statistical techniques: Regression, classification, and manifold learning. New York: Springer. Chapter 12

2. James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. New York: Springer Science + Business Media.

3. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning. New York: Springer Science + Business Media.

4. Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow. California: O'Reilly Media, Inc.

5. IBM. (2015/2016). IBM HR Analytics Employee Attrition & Performance [data file]. https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset