Predicting Employee Attrition

Kasey Moran

Abstract

Business Problem
Employee attrition is becoming a key problem in today's business landscape. Many companies are looking to understand what drives an employee to leave. Although some attrition is functional, a high level of attrition can cause a significant increase in hiring costs, employment costs, and training costs. Indirect costs also occur, such as increased workloads or decreased morale to existing employees taking departed employees' workloads. Companies want to see what they can change to the employee value proposition to retain key talent.
Expected Outcomes
From the dataset, I can evaluate the importance of factors from employees that have left the company. The variables will be used to create a model to predict if an employee will leave the company.

Data Exploration

Variable Name	Description
Age	Age
Attrition	If The Employee Left The Company
Business Travel	Frequency Of Business Travel
Daily Rate	Daily Rate
Department	Department Employee Works In
Distance From Home	The Distance From Work To Home
Education	Education Level
Education Field	Degree Field
Employee Count	Employee Count For Row
Employee Number	Employee Id Number
Enviroment Satisfaction	Satisfaction With Work Enviroment
Gender	Gender
Hourly Rate	Hourly Rate
Job Involvement	Level Of Job Involvement
Job Level	Level Of Job
Job Role	Job Role
Job Satisfaction	Satisfaction With Job
Marital Status	Maritial Status
Monthly Income	Monthly Income
Monthy Rate	Monthy Rate
Numcompanies Worked	Number Of Companies Worked At
Over 18	Is Age Over 18
Overtime	Whether Employee Works Overtime
Percent Salary Hike	Percentage Increase In Salary From 2017-2018
Performance Rating	Performance Rating
Relationship Satisfaction	Relationship Satisfaction
Standard Hours	Standard Hours
Stock Options Level	Level Of Stock Options
Total Working Years	Total Years Worked
Training Times Last Year	Hours Spent Training Last Year
Work Life Balance	Work Life Balance
Years At Company	Total Number Of Years At The Compnay
Years In Current Role	Years In Current Role
Years Since Last Promotion	Number Of Years Since Last Promotion
Years With Current Manager	Years Spent With Current Manager

Dataset:
https://www.kaggle.com/datasets/patelprashant/employee-attrition?select=WA_Fn-UseC_-HR-Employee-Attrition.csv

Data Pre-Processing

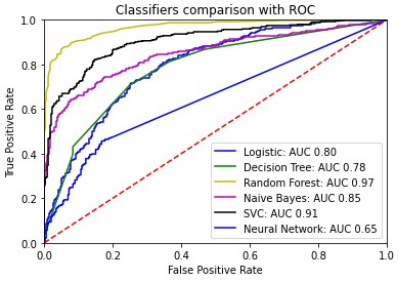
- Seven columns were excluded. EmployeeCount, EmployeeNumber, Over18, and StandardHours because the values were the same in all rows or irrelevant. I kept MonthlyIncome as the only variable related to salary and excluded the similar variables of DailyRate, HourlyRate and MonthlyRate
- Any rows with null values were dropped
- I converted the non-numeric variables Attrition, BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, and OverTime to dummy variables

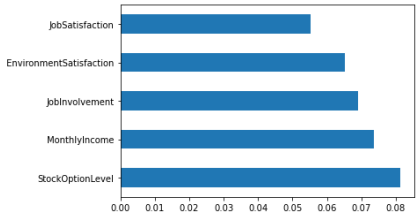
Model Creation

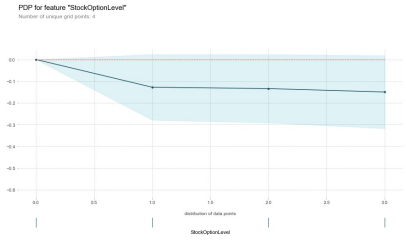
- X Variable Set: BusinessTravel, Department, Education, EducationField, EnvironmentSatisfaction, Gender, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MaritalStatus, MonthlyIncome, NumCompaniesWorked, OverTime, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager
- Y Variable: Attrition
- Training Data: 70% of total data
- Testing Data: 30% of total data
- SMOTE: Since the dataset was largely unbalanced with the minority class of Yes to Attrition, I applied SMOTE to address the imbalance
- Classification models evaluated: Logistic Regression, Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine, Neural Network

Model Evaluation

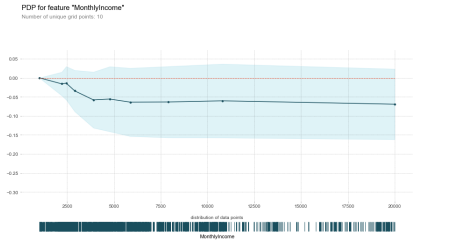
Model	Accuracy	Sensitivity	Specificity	Precision	F-1 Score
Logistic Regression	.7095	.6599	.7659	.76	.71
Decision Tree	.7270	.7056	.7514	.76	.73
Random Forest Max Features=5	.9068	.8706	.9480	.95	.91
Naïve Bayes	.7689	.7792	.7572	.79	.78
K-Nearest Neighbors	.7405	.8223	.6474	.73	.77
Neural Network	.4676	0	1	0	0
Support Vector Machine	.8216	.7868	.8613	.87	.82



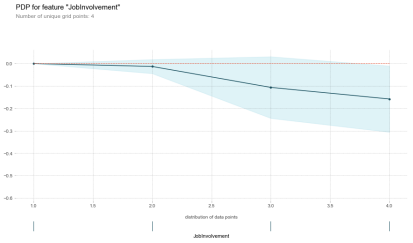




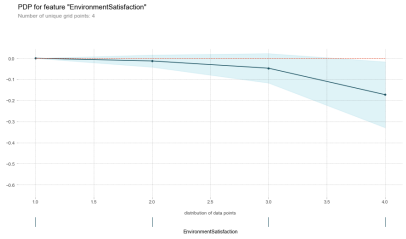
As shown in the plot above, there is an overall decrease in likelyhood of attrition with the increase in stock option level, especially from level 0 to 1. After level 1 the impact is not as strong.



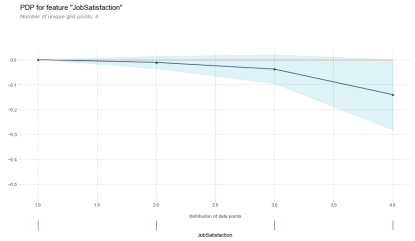
The increase in monthly income decreases the likelyhood of attrition overall with a slight peak at 2500 and 4800. The greatest impact occurs after the 2500 dollar mark up to about 5500 dollars. Prior to 2500 dollars and after 4000 dollars there is minimal impact of attrition.



There is a minimal decrease of attrition as job involvement increases from 1-2, then there is a significant decrease of attrition from the increased job involvement levels from 2 to 4.



Initially the environment satisfaction has a slight decrease in attrition as it increases from level 1 to 2. From 2 to 3 there is a larger decrease and from 3 to 4 there is a significant decrease in attrition.



Job satisfaction appears to decrease the likelihood of attrition, especially after the satisfaction passes a level of three.

Conclusion

Limitations and Improvements

- There are several variables with outliers, but they were not removed because there are many and they are very near each other
- To address the outliers, the data could be split to evaluate individual classes of variables. For example, splitting the data on the monthly income variable and looking at income above/below $16,000
- Some variables such as Job Satisfaction, Work Life Balance, Environment Satisfaction, Job Involvement and Relationship Satisfaction are subjective, and the meanings of levels could be different with different employees and vary over time
- As time passes, certain predictors can change due to economic events, cultural shifts etc. Generalization without time data may not be dependable
- Since there was a small number of employees with attrition, the dataset required synthetic oversampling

Conclusion

From the analysis of employee data, I was able to develop a model for predicting employee attrition. This would be advantageous to companies looking to retain talent and decrease the costs associated with attrition.
The random forest model with five features proved to show the best performance between the models. The strongest predictors were Stock Option Level, Monthly Income, Job Involvement, Environment Satisfaction and Job Satisfaction, respectively. Of those predictors all decreased the likelihood of attrition overall with an increase in level.