HR Analytics and Prediction of Employee Attrition

Kesav Adithya Venkidusamy
DSC680-T301 Applied Data Science, Bellevue University
Professor Catherine Wiliams

Summary


Introduction
Data Mining & EDA
Predictive Modeling & Analysis
Findings & Recommendation
Ethical Considerations

Employee Attrition Prediction Life Cycle

Introduction

Employee retention strategies are integral to the success and well-being of a company. Attrition is a problem that impacts all businesses, irrespective of geography, industry and size of the company.

Employee attrition leads to significant costs for a business, including the cost of business disruption, hiring new staff and training new staff.

There are often many reasons why employees leave an organization, and in this case study, I will explore some of the key drivers of employee attrition

Data Mining - Data Structure

Feature Name	Feature Description	Feature Type
Attrition	Person has left the company or not	Target
Age	Age of the person	Continuous
BusinessTravel	How frequently the person travels	Discrete
DailyRate	Daily Rate for the employee	Continuous
Department	Department of the person	Discrete
DistanceFromHome	Distance of the company from home	Continuous
Education	Education of the person	Discrete
EducationField	Education field of the person	Discrete
EmployeeCount	Count of Employee	Discrete
EmployeeNumber	Employee id of the person	Continuous
EnvironmentSatisfaction	Environment Satisfaction	Discrete
Gender	Gender	Discrete
HourlyRate	Hourly rate for the employee	Continuous
JobInvolvement	Involvement in Job	Discrete
JobLevel	Job Level	Discrete

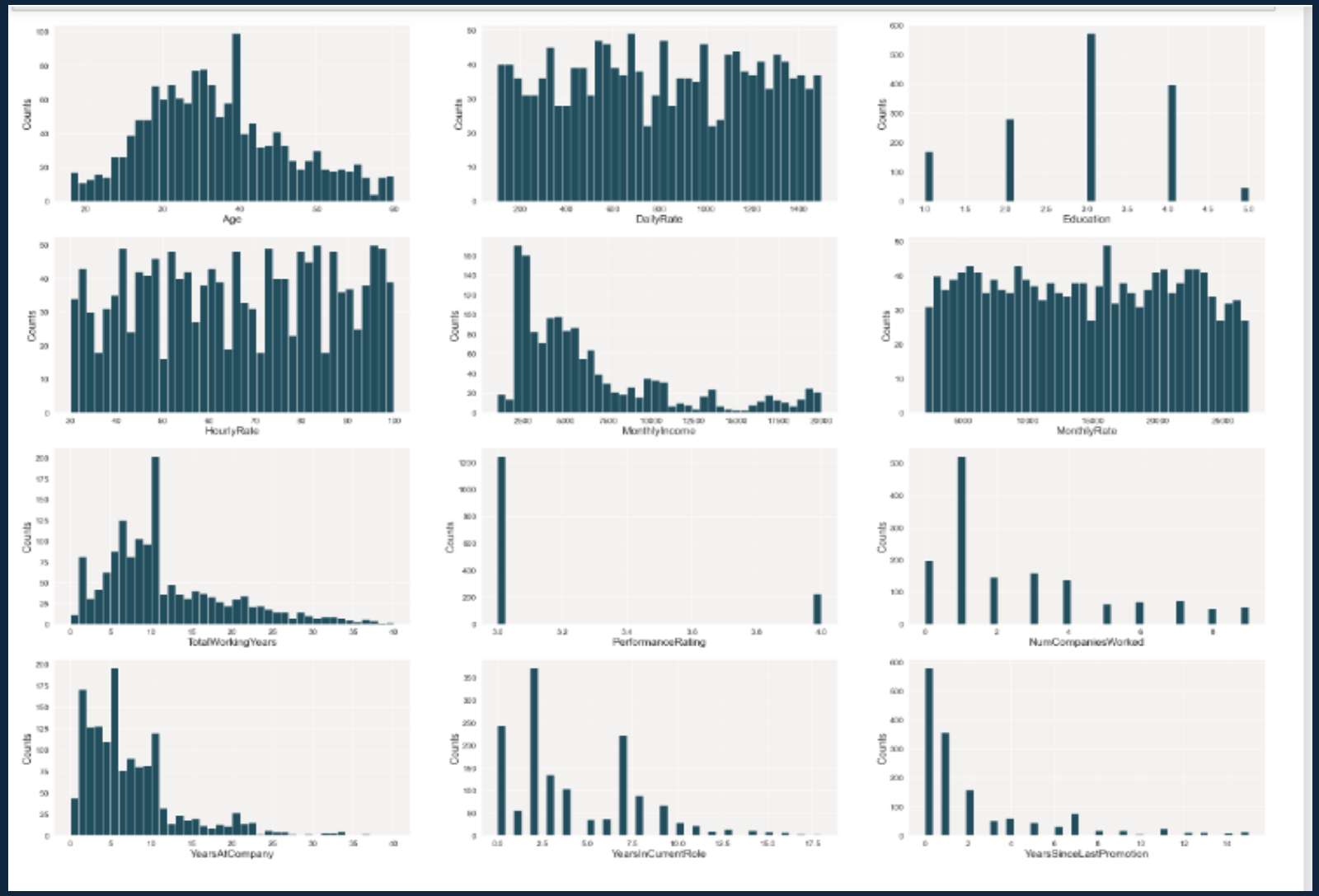
Features	Feature Description	Feature Type
JobRole	Job Role	Discrete
JobSatisfaction	Job Satisfaction	Discrete
MaritalStatus	Marital Status of Employee	Discrete
MonthlyIncome	Monthly Income of the person	Continuous
MonthlyRate	Monthly Rate	Continuous
NumCompaniesWorked	Number of Companies worked	Discrete
Over18	Over 18 years	Discrete
OverTime	Worked over time	Discrete
PercentSalaryHike	Percentage of Salary Hike	Continuous
PerformanceRating	Performance Rating	Discrete
RelationshipSatisfaction	Relationship Satisfaction for the employee	Discrete
StandardHours	Standard work hours	Discrete
StockOptionLevel	Stock Option Level given to the employee	Discrete
TotalWorkingYears	Total Number of years worked	Continuous
TrainingTimesLastYear	Training times attended during last year	Continuous
WorkLifeBalance	Work Life Balance	Discrete
YearsAtCompany	Number of years with current company	Continuous
YearsInCurrentRole	Number of years in the current role	Continuous
YearsSinceLastPromotion	Number of years since last promotion	Continuous
YearsWithCurrManager	Number of years with current manager	Continuous

Data Cleaning

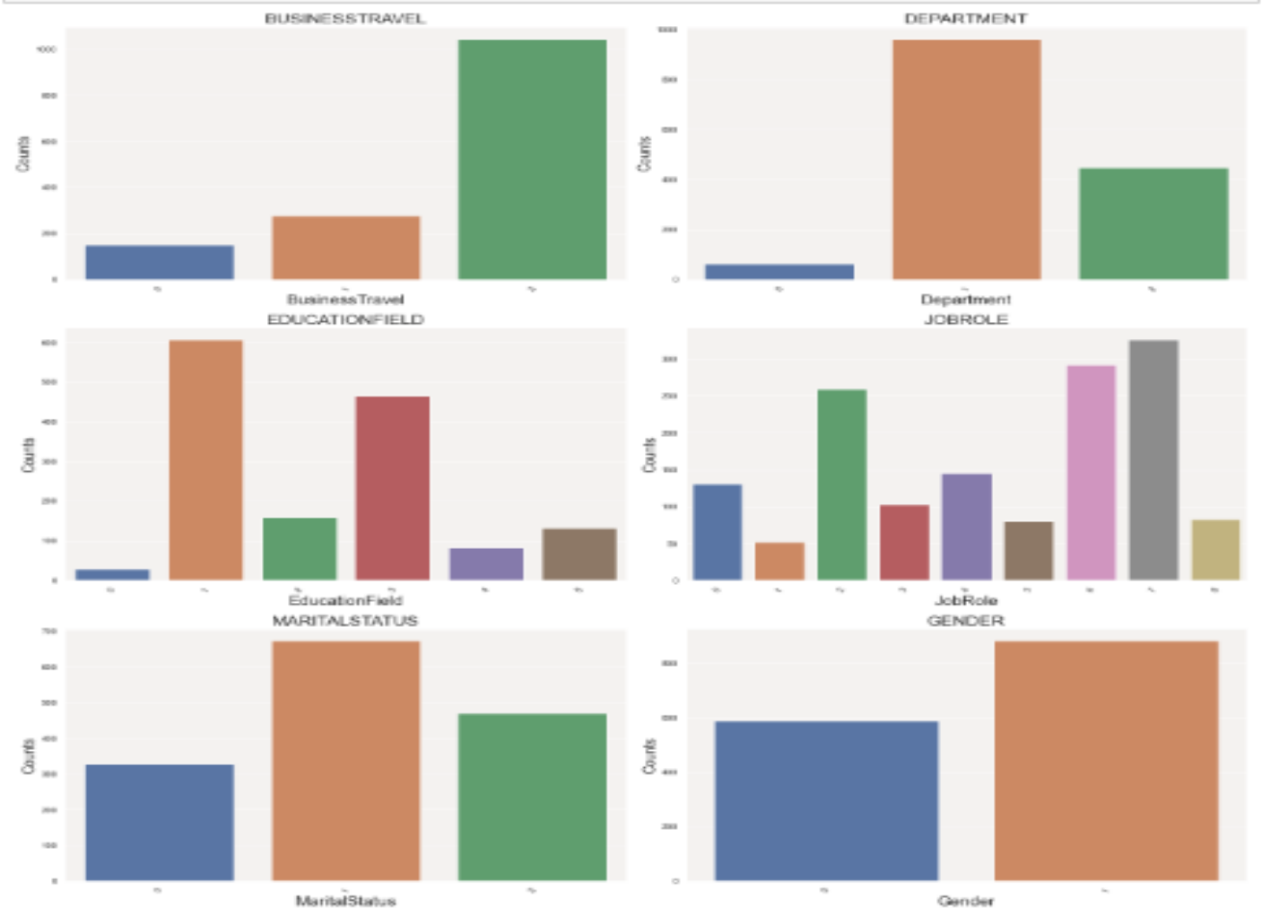
- Missing/Null Check - No missing value present in any of the feature
- Duplicate Check - No duplicate value in the dataset
- Unwanted features Removal – The below irrelevant features are removed from the dataset
  - EmployeeCount - constant value "1“
  - StandardHours - constant value "80“
  - Over18 - constant value "True“
  - EmployeeNumber - Employee number is key column which can also be removed
  - StockOptionLevel - I believe this is stock options given to the employees having values between 1 to 3;

EDA - Numerical Variables

- Age
- Daily Rate
- Education
- Hourly Rate
- Monthly Income
- Monthly Rate
- TotalWorkingYears
- PerformanceRating
- NumCompaniesWorked
- YearsAtCompany
- YearsInCurrentRole
- YearsSinceLastPromotion

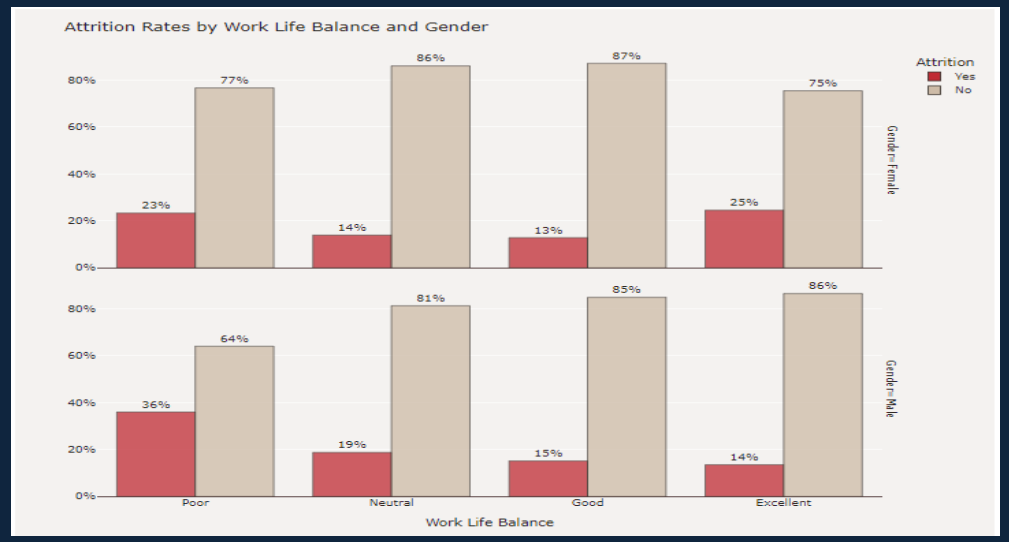


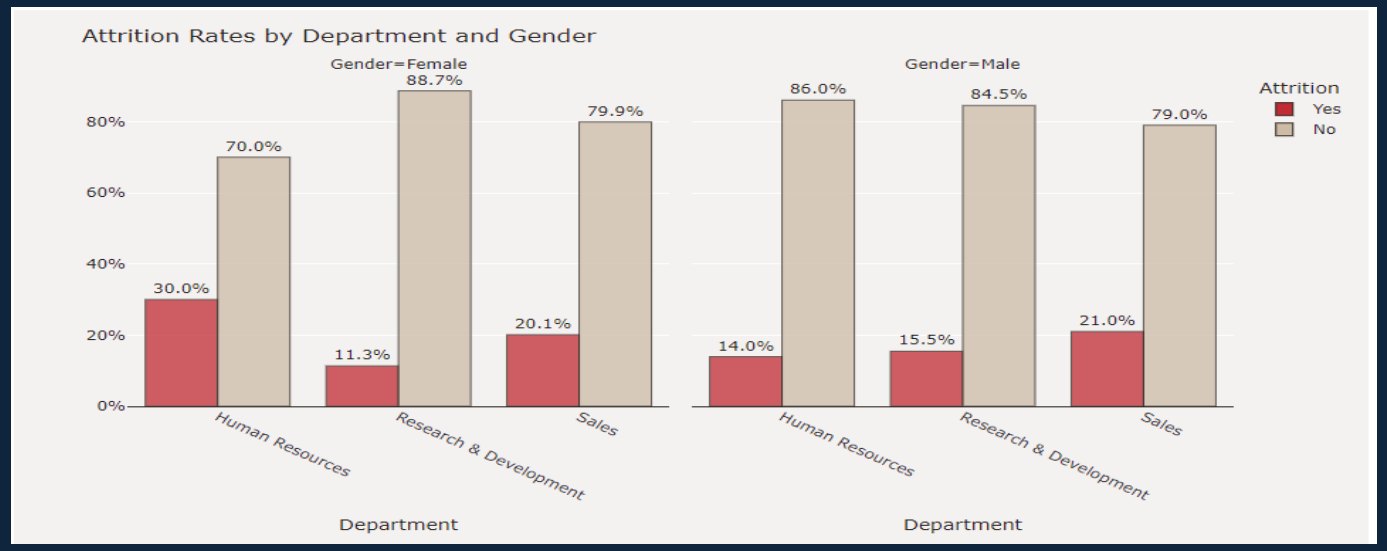
EDA - Categorical Variables

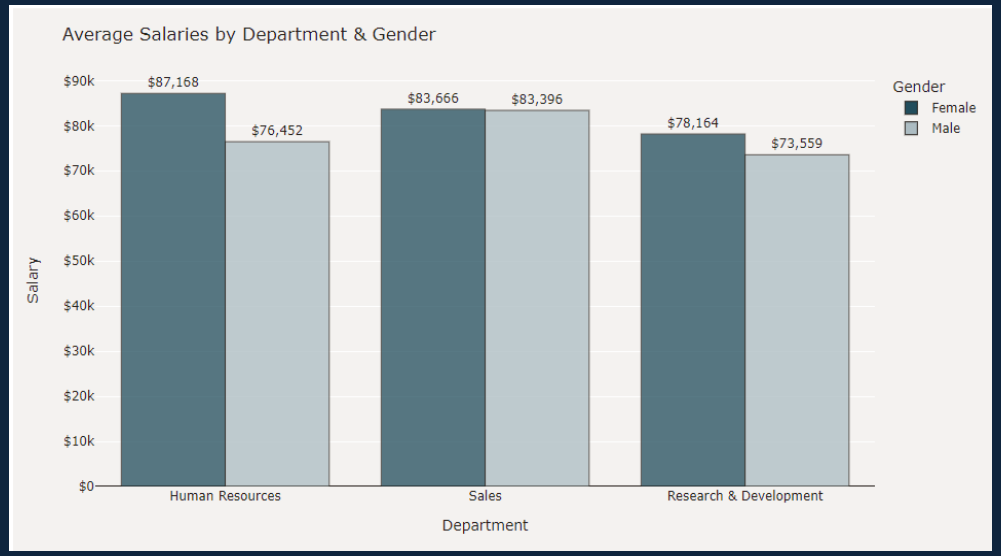


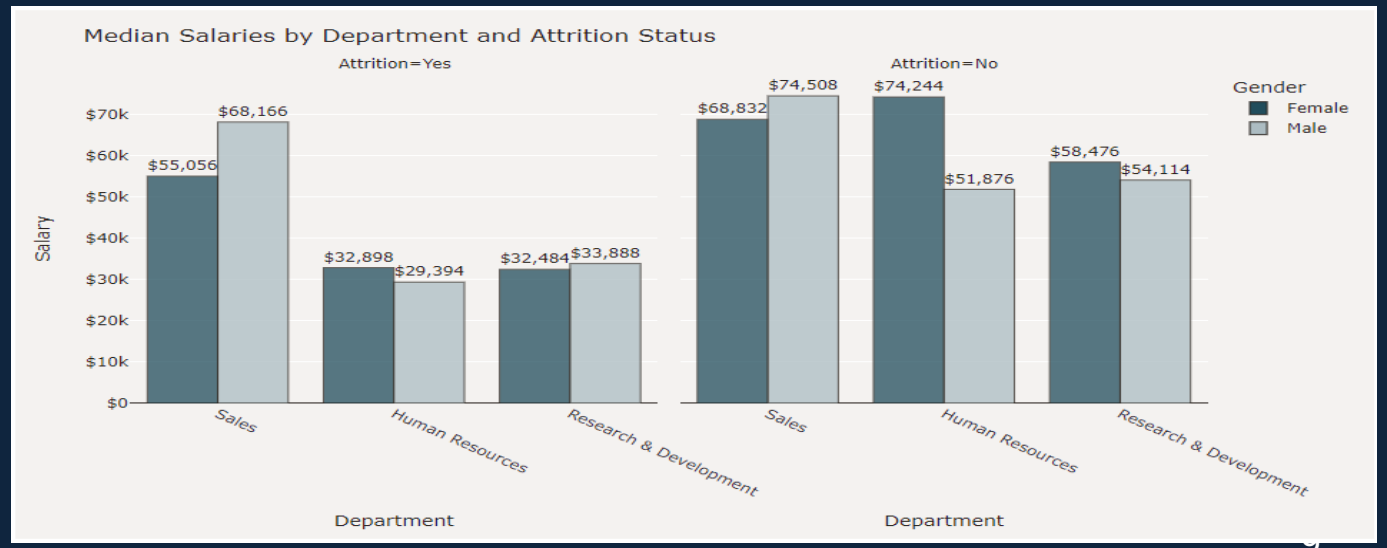
- Business Travel
- Department
- EducationField
- JobRole
- MaritalStatus
- Gender

EDA - Target Variable “Attrition” Analysis

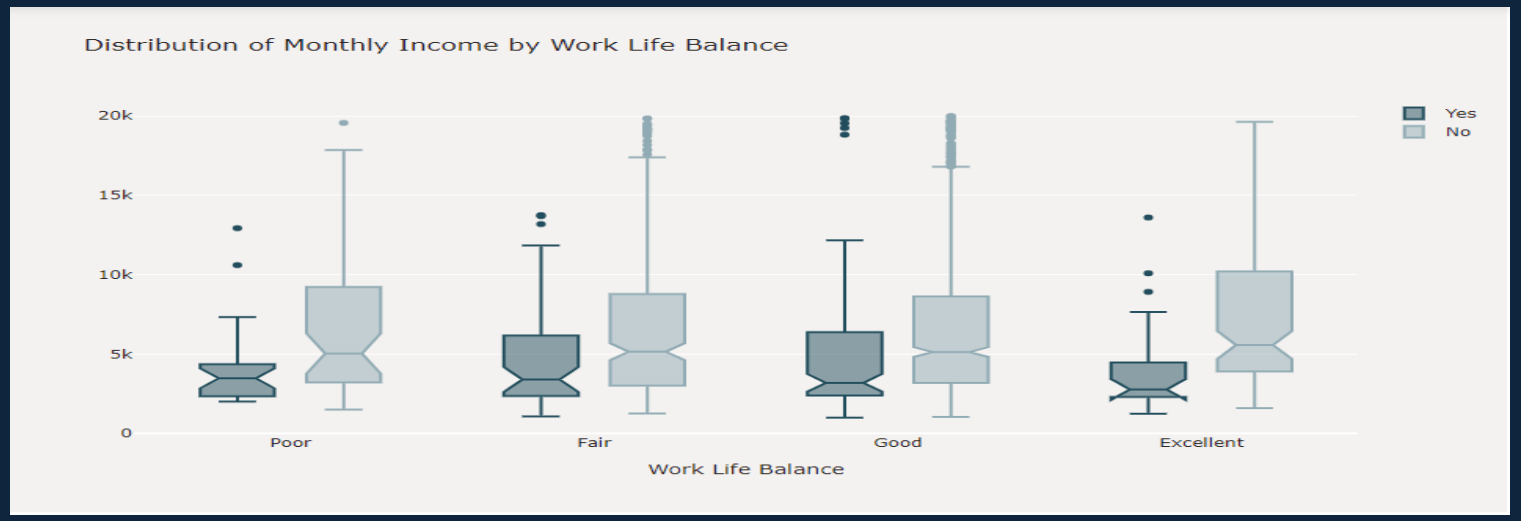






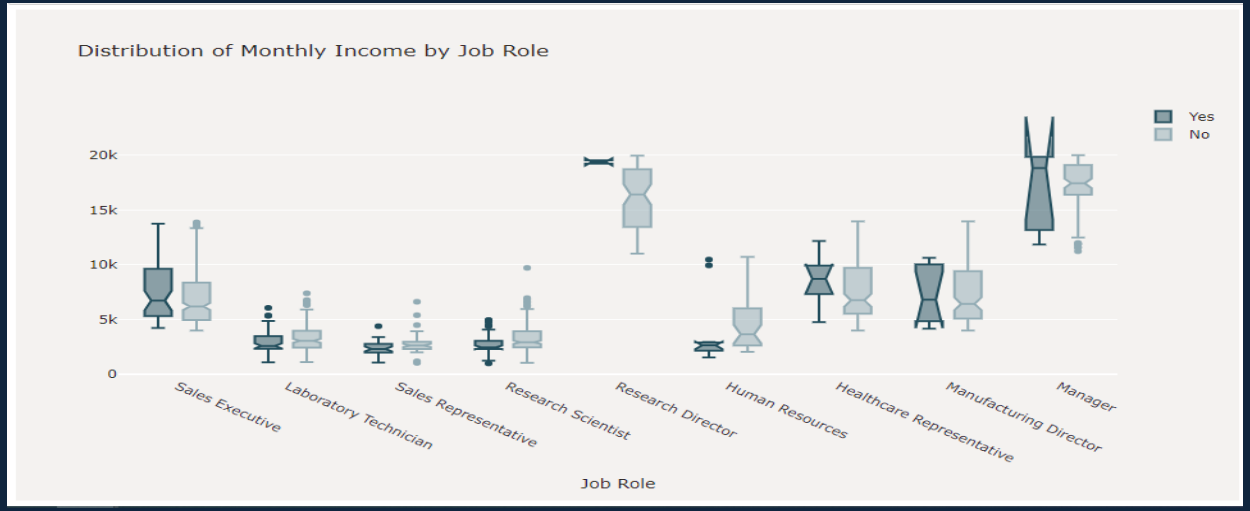


EDA - Distribution of Monthly Income by Work Life Balance and Job Role

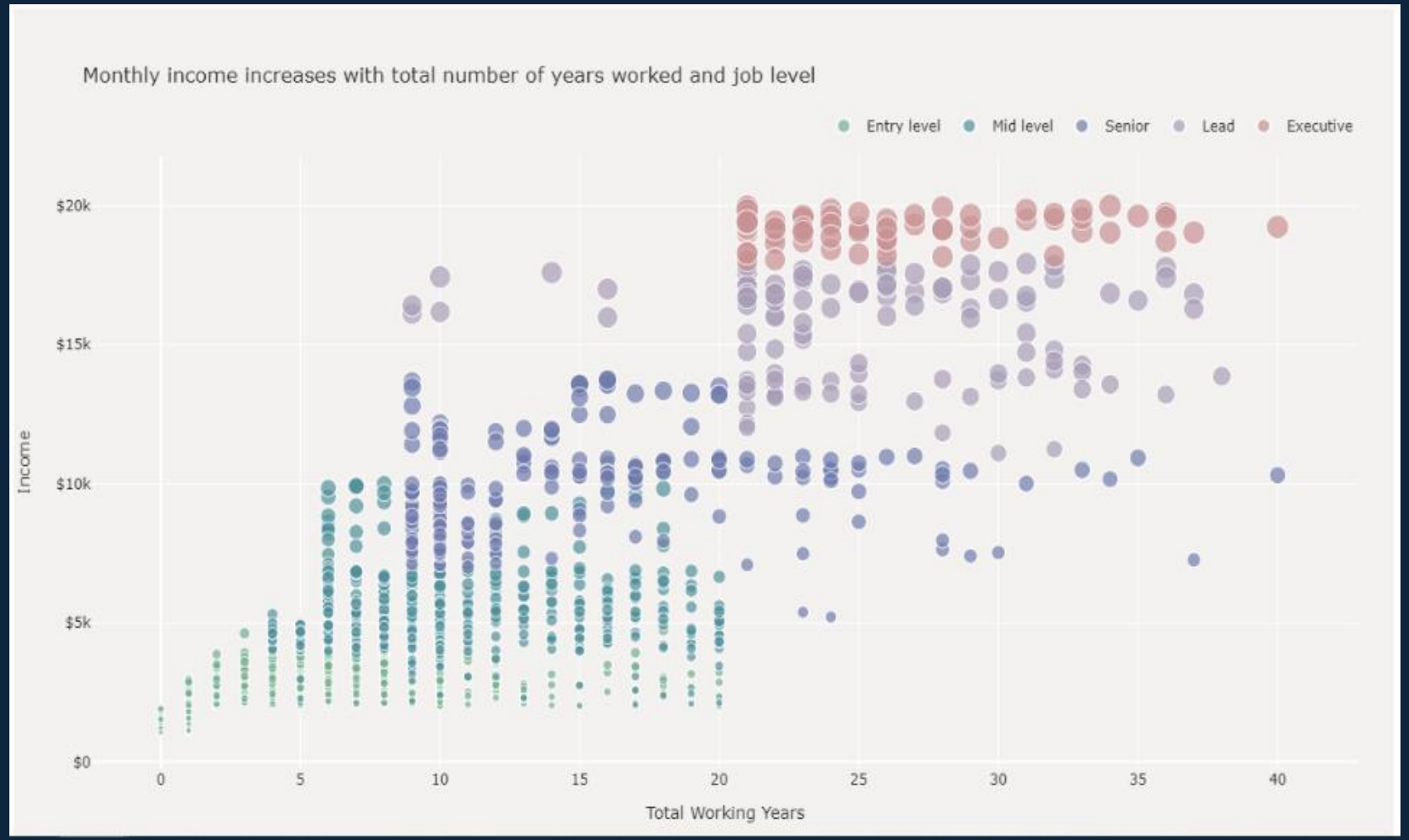


The Attrition rate is high among the people who have "Good" work life balance earning median monthly income of 3202 dollars. However, the attrition rate is less among the people who have "Excellent" work life balance earning median monthly income of 2785 dollars; The people with "Fair" work life balance comes next.

The attrition rate is high among "Managers", "Manufacturing Director" and "Sales Executives" whereas it is low for "Research Director" and "Sales Representative"; So, "Research Director" and "Sales Representative" are not willing to quit the company often whereas "Managers", "Manufacturing Director" and "Sales Executives" often change their companies.



EDA - Distribution of Monthly Income by Work Life Balance and Job Role



Monthly income is positively correlated with total number of years worked and there is strong association between an employee's earnings and their job level.

Feature Engineering

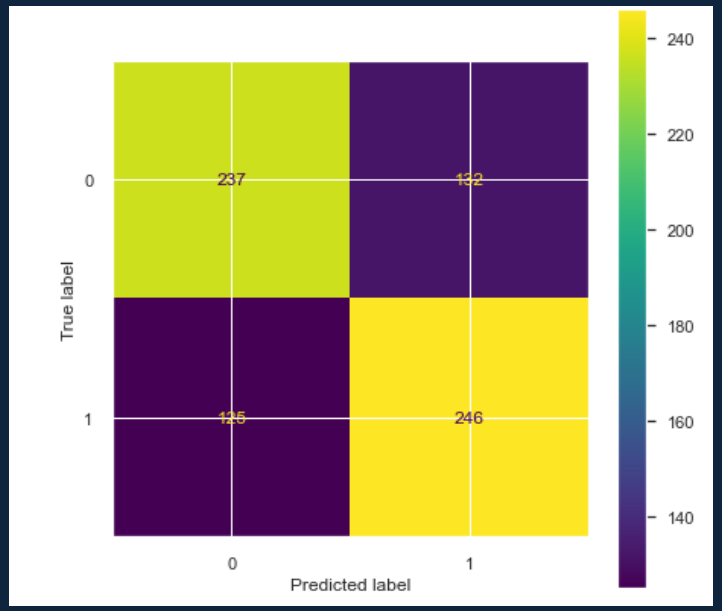
The below methods were applied on the features prepping them for model building

- Label Encoder
- SMOTE

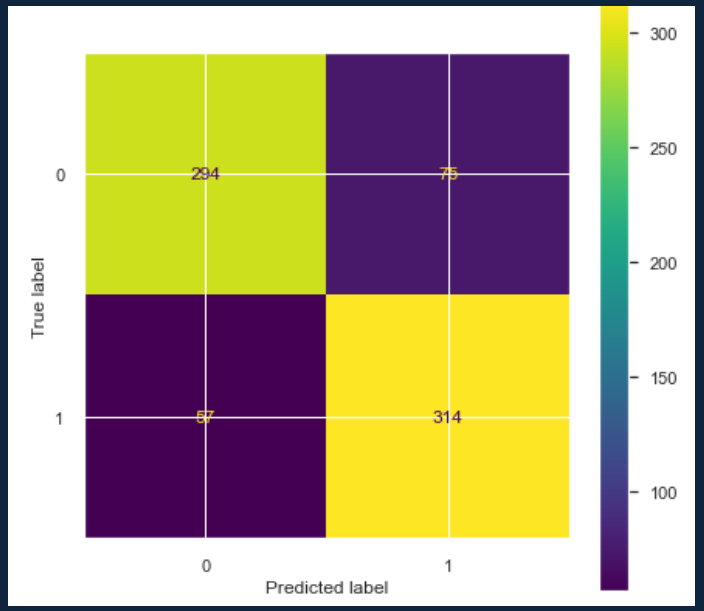
Predictive Modeling

Model	Definition	Pros	Cons
Logistic Regression	Predicts a dependent data variable by analyzing the relationship between one or more existing independent variables	Probabilistic Approach, gives info about statistical significance of features	the assumption of linearity between the dependent variable and the independent variables
Decision Tree	A series of sequential decisions made to reach a specific result	Interpretability, no need for feature scaling, works on both linear/nonlinear problems	Poor results on small datasets, overfitting can easily occur
Random Forest	A forest of randomly created decision trees, a combined output of individual decision trees to generate the final output.	Powerful and accurate, good performance on many problems including nonlinear	No interpretability, overfitting can easily occur, need to choose the number of trees

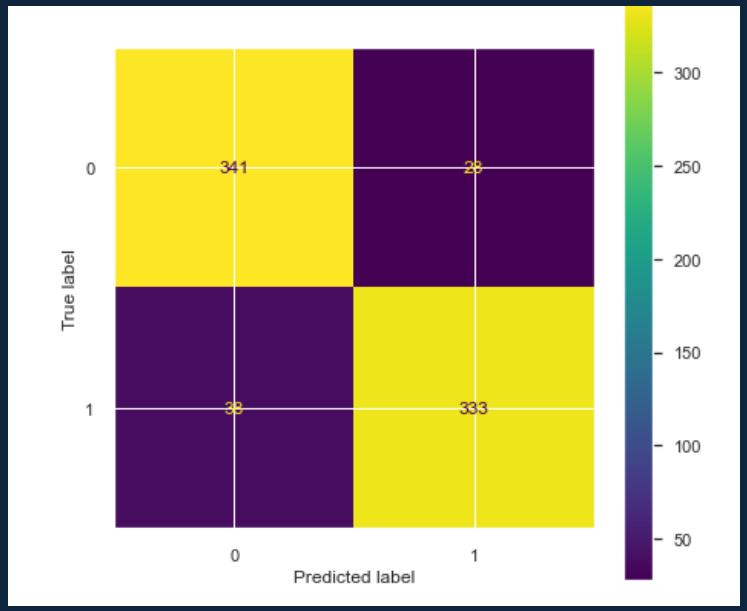
Confusion Matrix



Logistic regression



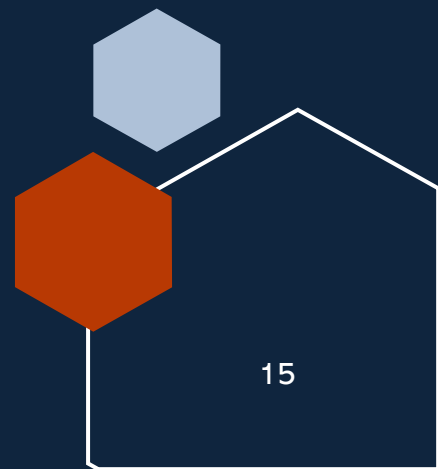
Decision Tree



Random Forest

Model Efficiency Measures

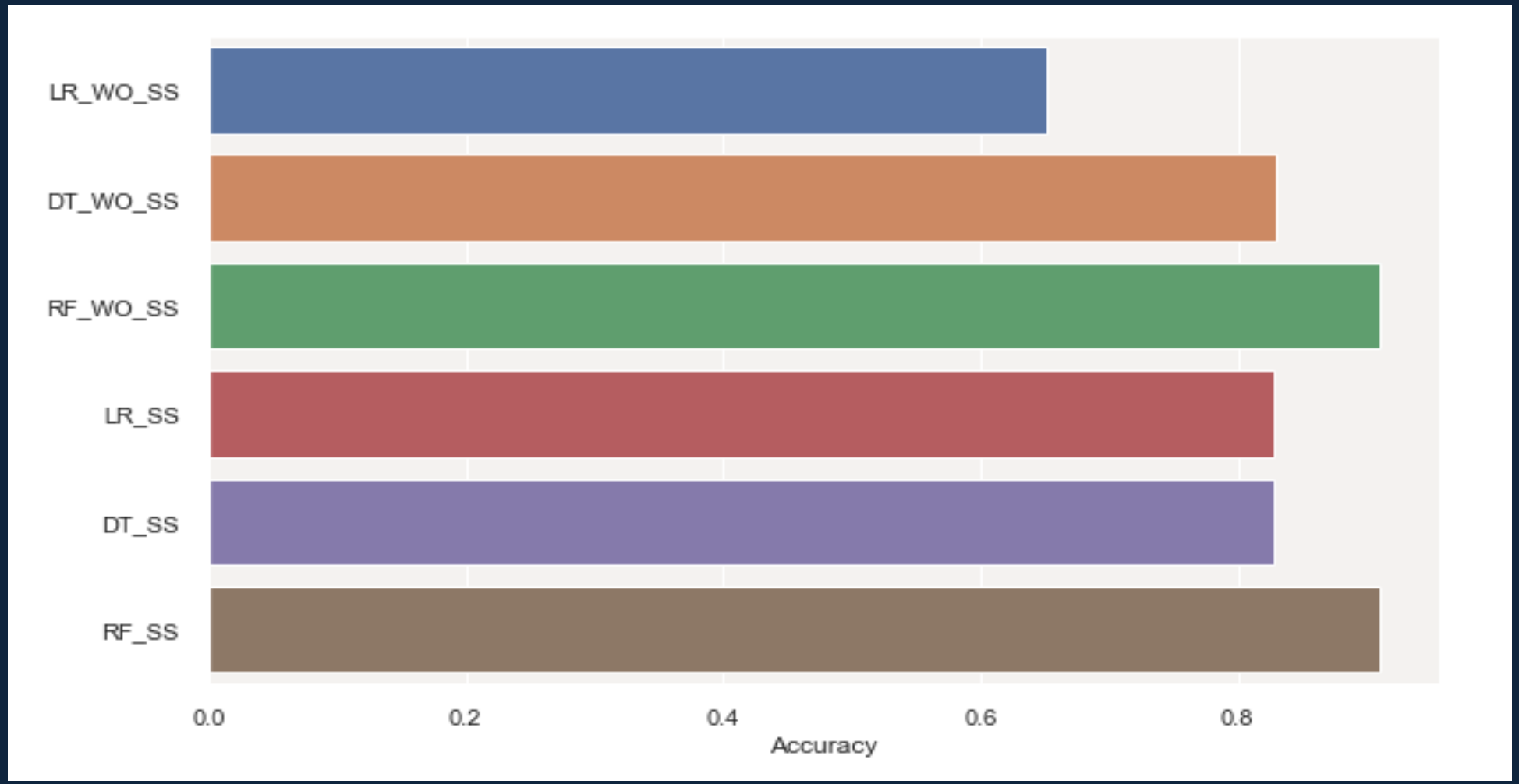
- **Accuracy** - Accuracy represents the number of correctly classified data instance over the total number of data instances
- **F1 Score:** F1-Score is a metric which takes both precision and recall into account.
  - **Precision:** Positive predictive value
  - **Recall:** true positive rate
- **AUC Score:** What area under the ROC curve describes good discrimination? We will use the following rule of thumb
  - 0.5: This suggests no discrimination, so we might as well flip coin
  - 0.5-0.7: We consider this as poor discrimination, not much better than a coin toss
  - 0.7-0.8: Acceptable discrimination
  - 0.8-0.9: Excellent discrimination
  - >0.9: Outstanding discrimination



Model Efficiency

Model	Standard Scalar	Accuracy	F1 Score (Income <=50K)	F1 Score (Income >50K)	AUC Score
Logistic Regression	No	65.27%	0.65	0.66	0.71
	Yes	81.62%	0.83	0.82	
Decision Tree	No	82.16%	0.83	0.82	0.83
	Yes	82.16%	0.83	0.82	
Random Forest	No	91.08%	0.91	0.91	0.97
	Yes	90.81%	0.91	0.91	

Model Efficiency – Score Chart



Analysis - Best Performing Features

Pearson's correlation matrix Feature correlation to target variable “Attrition"
Chi-Squared (X2) Test 5 Best features correlated to “Attrition"
Using Feature Importance of Random Forest Classifier

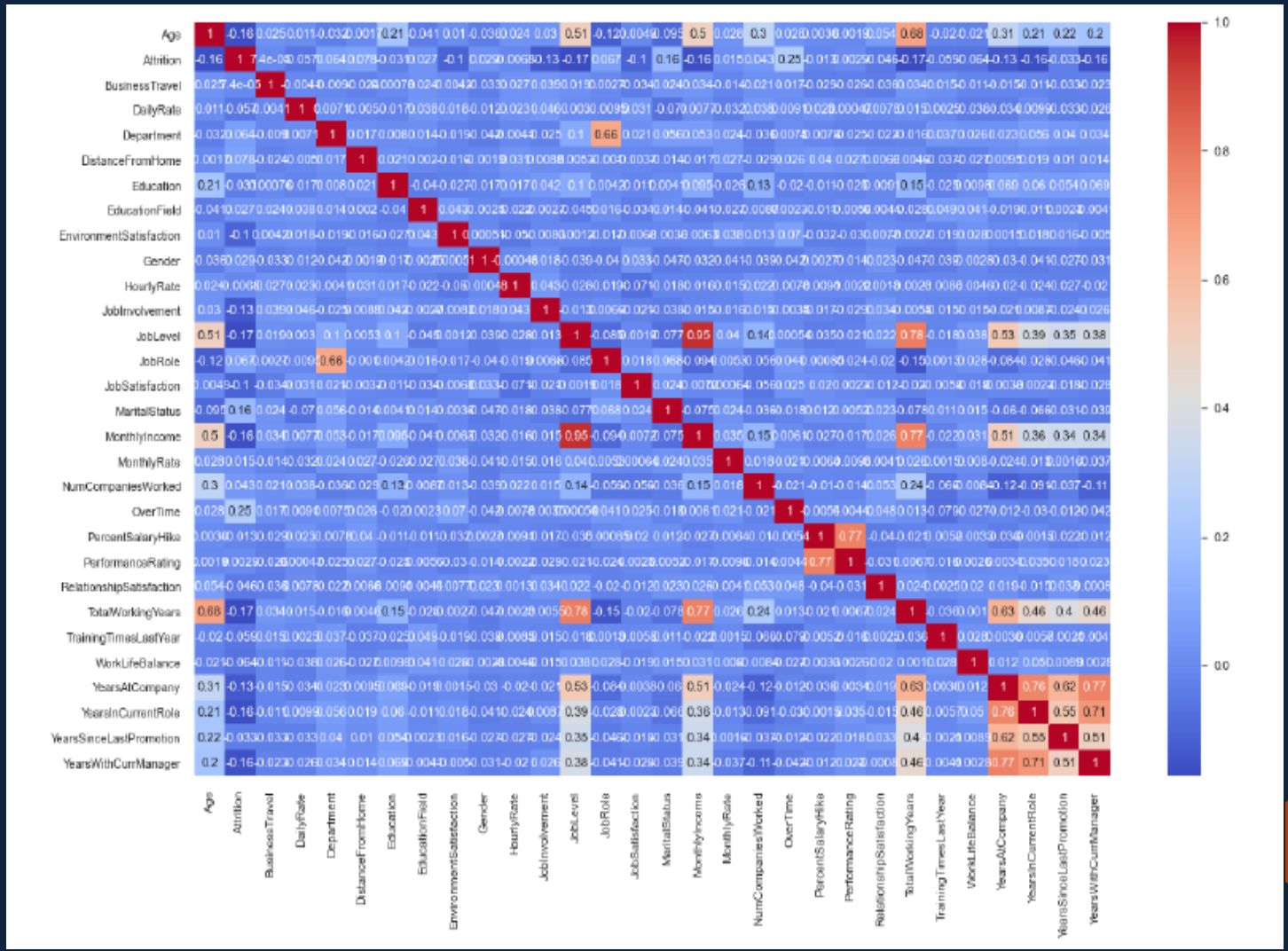
Pearson’s Correlation Matrix

Positive Correlation:

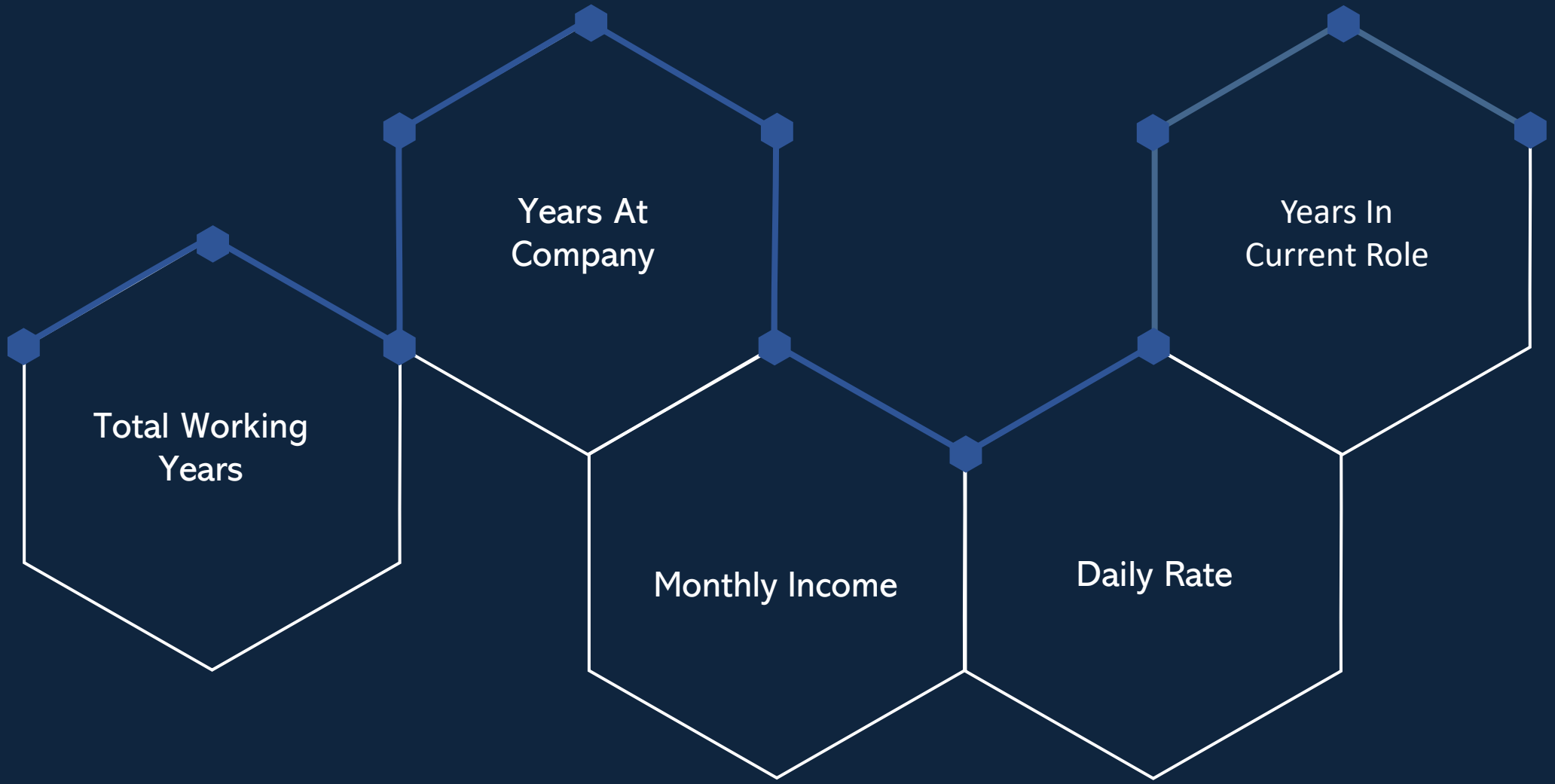
- OverTime
- MaritalStatus
- DistanceFromHome

Negative Correlation:

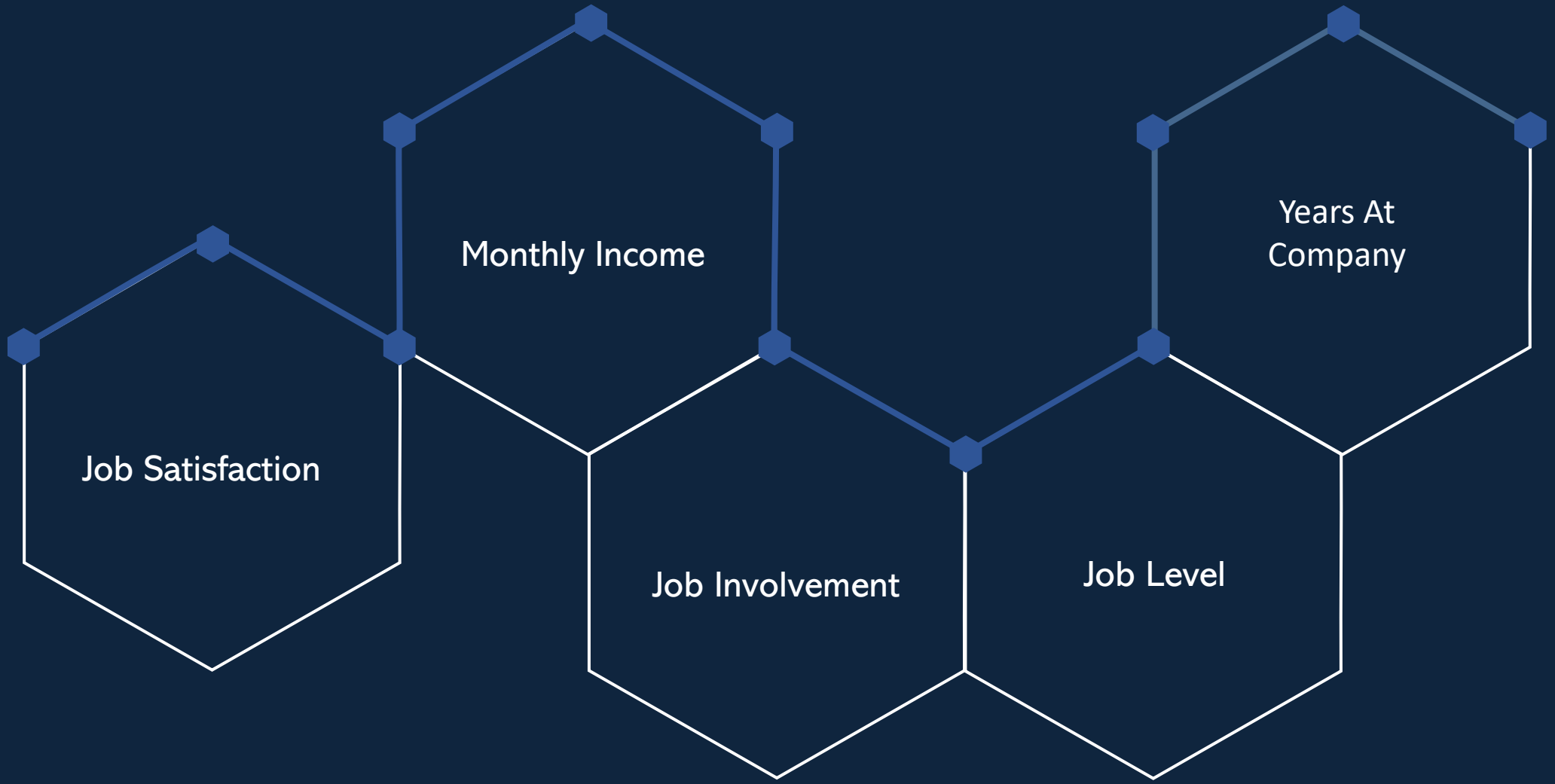
- TotalWorkingYears
- YearsInCurrentRole
- MonthlyIncome



Chi-Squared (X2) Test - 5 Best features



Feature Importance of Random Forest Classifier



Findings and Recommendation

Model

- Random Forest Classifier is the best model to predict the employee attrition

Features having high impact on the target variable “Attrition”

- Job Satisfaction - Job Satisfaction of the employees
- Monthly Income - Monthly Income earned by the employees
- Years At Company - Employee’s experience
- Over time - Over time
- Years in Current Role - Number of years in current role

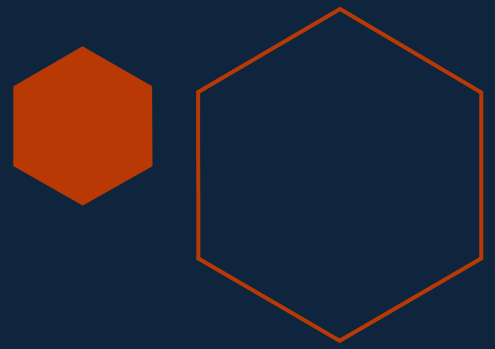
Ethical Considerations

- Consideration of result from the analysis in decision making. Some of the conclusions make from this project's study could be incorrect or misrepresented due to insufficient or incorrect data. So, while sharing the outcome of this project to larger audience, the underlying assumptions and data considerations should be shared.

- No personal and sensitive information is used in the dataset. . Since this dataset is fictional created by IBM data scientists, this is already taken care by them and personal identifying information (like gender, age) is broad enough which is untraceable to any individual.

References

- Pavansubhash (2017). IBM HR Analytics Employee Attrition & Performance. https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset
- Emily Killham (January 25, 2022). Employee Attrition Analytics: The Who, When & Why Of Employee Turnover. https://blog.perceptyx.com/employee-attrition-analytics
- Maggie Wooll (January 24, 2022). Fighting employee attrition: What is within your control? https://www.betterup.com/blog/employee-attrition
- Unites States Depart of Labor. Annual quits rates by industry and region, not seasonally adjusted. https://www.bls.gov/news.release/jolts.t18.htm






Thank you