**Q1. Data Preprocessing.**

We have 35 features for this dataset and 1470 data points. The data set is about employee attrition.

For the following features we had to convert them from numerical to categorical for their

variable type transformation:

- Job Level
- Stock Option Level
- Job Satisfaction
- Job Involvement
- Environment Satisfaction
- Education
- Performance Rating
- Relationship Satisfaction
- Work Life Balance

The following features were excluded as they have no variance and do not contribute towards the outcome of the analysis:

- Employee Count
- Over 18
- Standard Hours

A new feature list with 31 features with the transformed ones and after excluding the ones with no variance, was created.


| Feature Name | Data Quality | Index | Importance | Var Type | Unique | Missing | Mean | Std Dev | Median | Min | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Attrition | | 2 | Target | Categorical | 2 | 0 | | | | | |
| TotalWorkingYears | | 29 | | Numeric | 39 | 0 | 11.15 | 7.69 | 10 | 0 | 40 |
| YearsAtCompany | | 32 | | Numeric | 36 | 0 | 6.99 | 6.06 | 5 | 0 | 40 |
| YearsInCurrentRole | | 33 | | Numeric | 19 | 0 | 4.22 | 3.61 | 3 | 0 | 18 |
| OverTime | | 23 | | Categorical | 2 | 0 | | | | | |
| JobRole | | 16 | | Categorical | 9 | 0 | | | | | |
| YearsWithCurrManager | | 35 | | Numeric | 18 | 0 | 4.14 | 3.55 | 3 | 0 | 17 |
| Age | | 1 | | Numeric | 43 | 0 | 36.91 | 9.17 | 36 | 18 | 60 |
| JobLevel_Categorical | | 15 | | Categorical | 5 | 0 | | | | | |
| MonthlyIncome | | 19 | | Numeric | 1,096 | 0 | 6,491 | 4,745 | 4,933 | 1,009 | 19,999 |

**Q2. Business Case.**

For any organization, the cost of rehiring another person when an employee leaves the firm is highly expensive due to various factors. It is better for the organizations that the employees can continue to work for them. The goal is to reduce employee attrition so that the cost of rehiring and training the next person can be saved.

We will use a regression model, random forest, boosted trees, KNN, neural networks, SVM to analyze the data. The data that we are considering 31 features and 1470 data points.

Financial implications:

| | | |
|---|---|---|
| Cost saved per employee | $65,000 | 10 months have been assumed as Cost of Retirement, Training and Lost Productivity |
| Cost after Employee leaves | $45,500 | 7 months have been assumed as the cost of rehiring and productivity loss has been considered |

The mean for the monthly salary in the dataset is $6250. So, the annual income is $78,000.

| Retention cost | |
|---|---|
| Bonus provided | $11,700 |
| Events for the team | $1000 |
| Other strategies | $1000 |
| Total Cost incurred | $13,700 |

*15% on $78,000 (annual income as per the mean)*

To keep the employees we provide a 15% bonus on the annual income. We would also try to keep events that make the employee more belonging and connected to the team and the organization. There can also be other retention strategies which will be applied so we will keep a budget of $1,000 aside for that as well.

**Payoff Matrix**

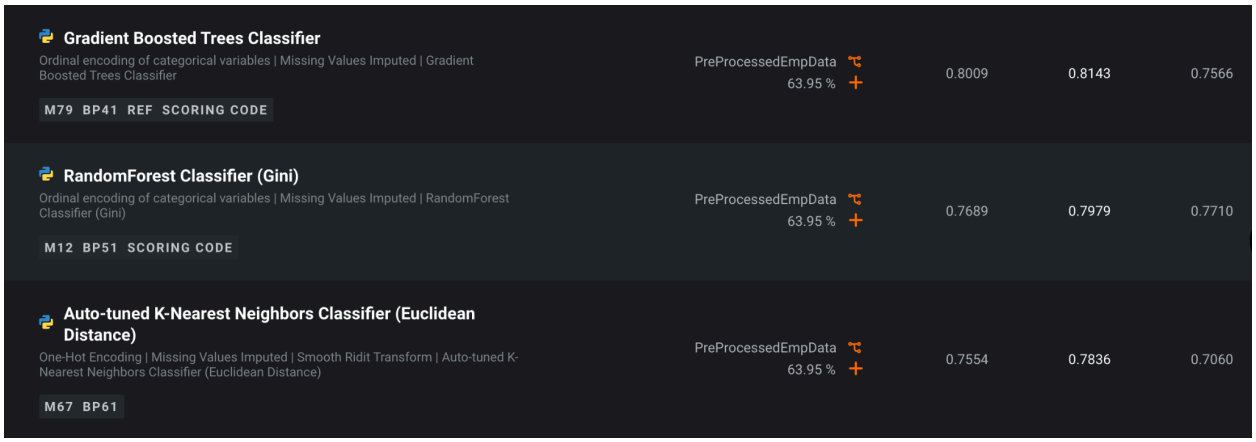
| | |
|---|---|
| TP | The model says employee will leave, and we provide them with bonus and implement retention strategies so the cost will be (considering all employees provided with bonus stay with the company)<br>= Cost saved - retention cost<br>= $65,000 - $13,700<br>= $51,300 |
| TN | The model predicts the employee will not leave and the employee doesn't leave but we still spend around $2,000 to keep the employee engaged and involved with the company. |
| FP | The employee wouldn't leave but the model falsely predicts that the employee would leave, so we end up giving bonus and implement the retention strategies for them = $13,700 |
| FN | The model predicts the employee wouldn't go but they leave, so the cost incurred is = $45,500. |

| | Predicted | | |
|---|---|---|---|
| Actual | | 0 | 1 |
| | 0 | TN (-2000) | FP (-13700) |
| | 1 | FN (-45500) | TP (51,300) |

**Q3. Models and metrics.**

Our target variable is attrition as we want to understand whether the employee will stay with the company or leave the company.

The models we have considered here are logistic regression, random forest,  boosted trees, KNN, neural networks, and SVM.

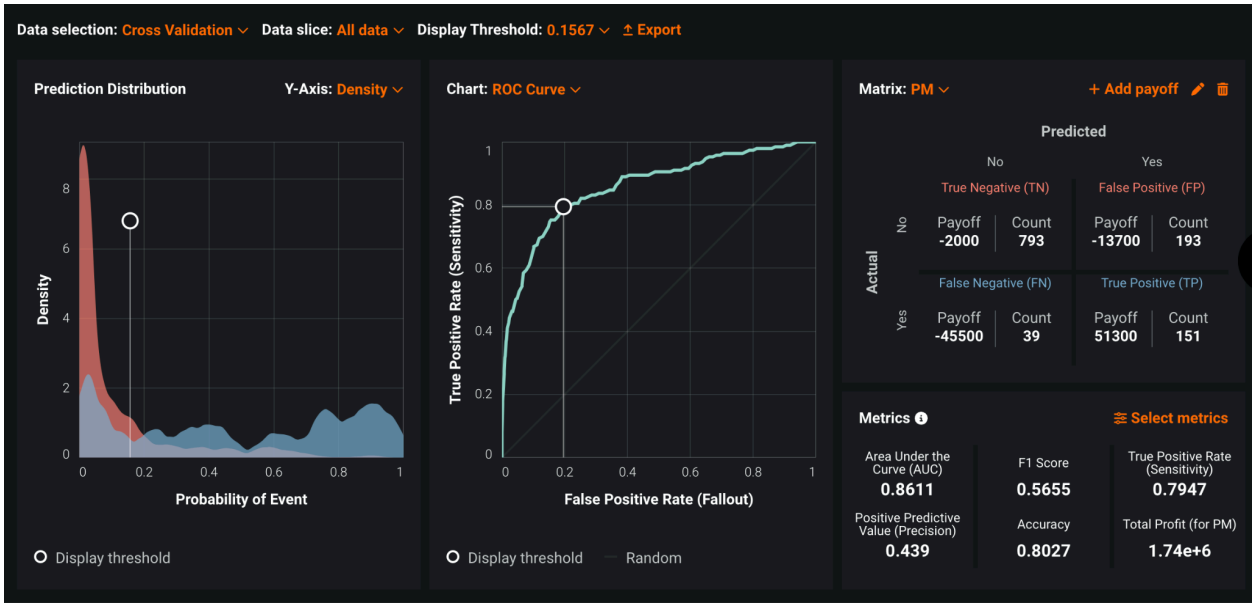




The values for AUC look good for both cross validation and holdout. It can be seen that considering AUC, Neural Network Classifier performs the best here, and Logistic Regression has a very close value of AUC to it too.

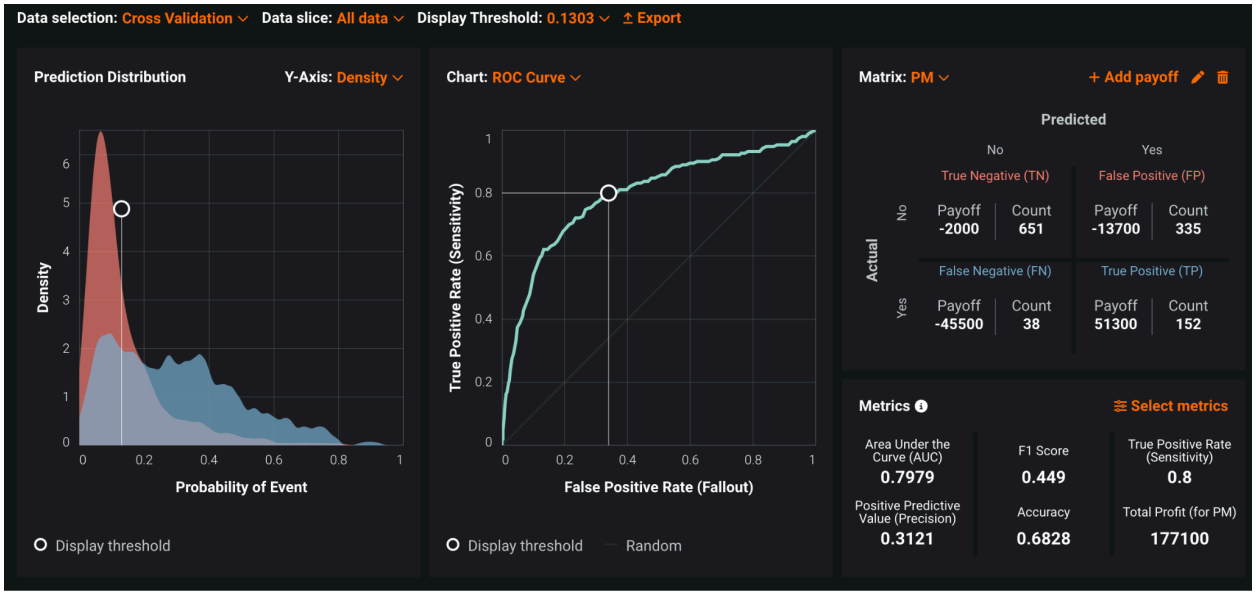
**Cross Validation comparison of all models:**

All the values are captured by keeping the maximum payoff.

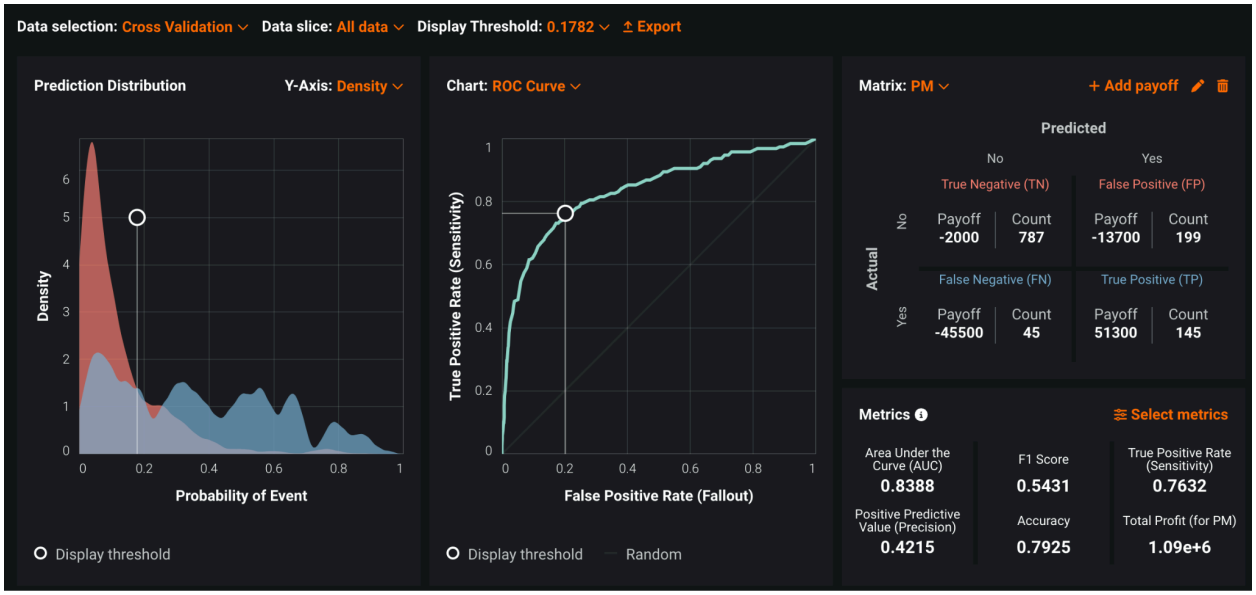
### *Logistic Regression*



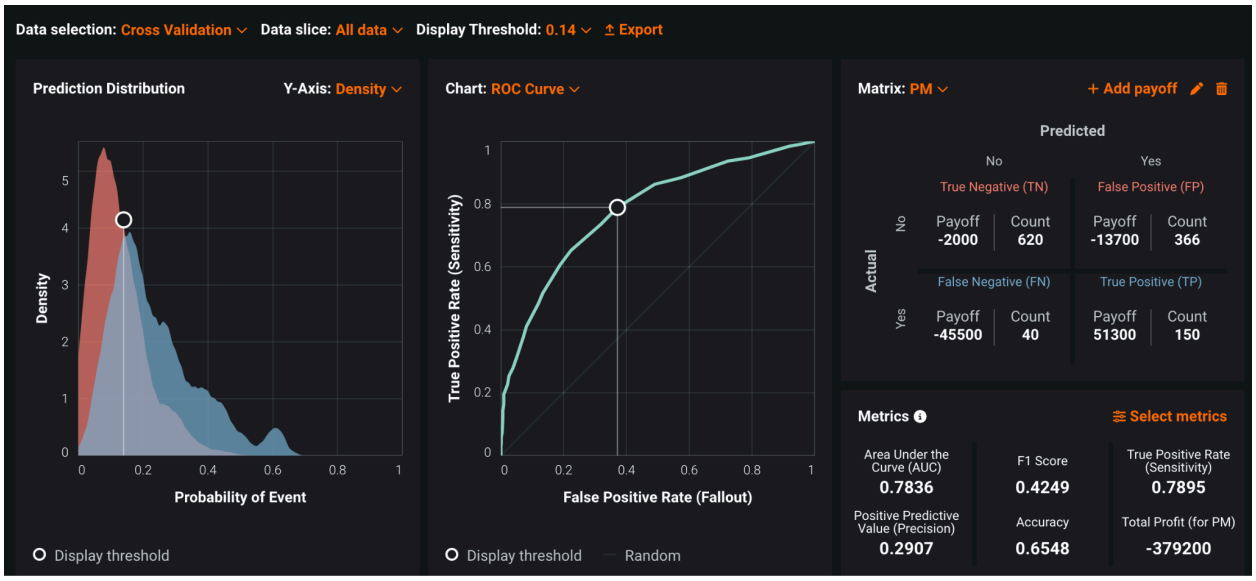

### *Random Forest Classifier*



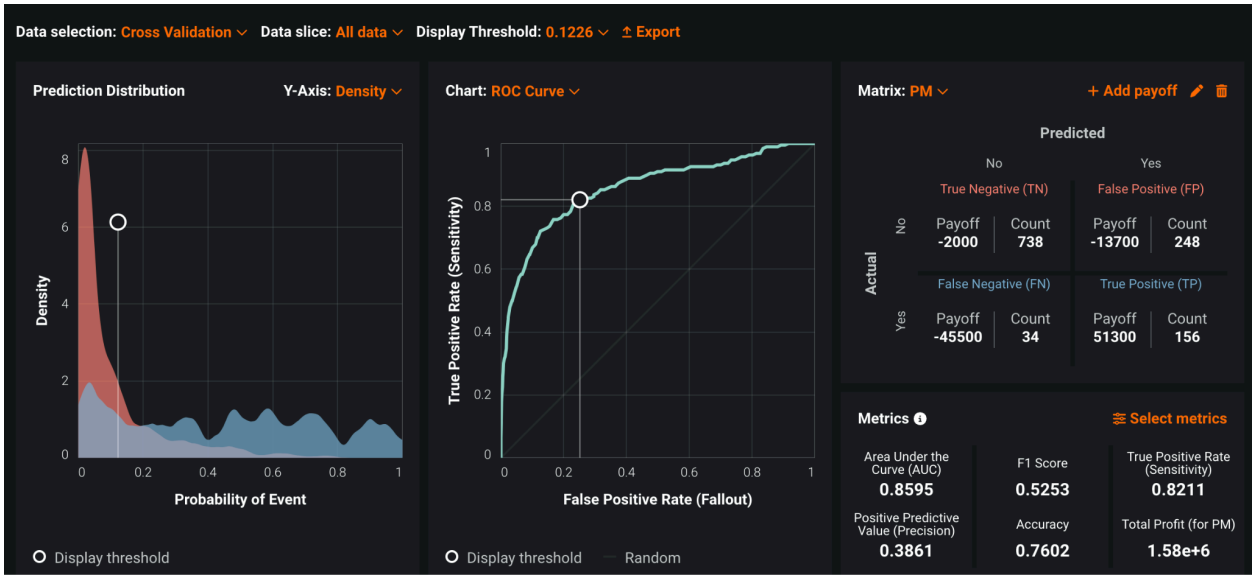
## *Boosted Trees Classifier*



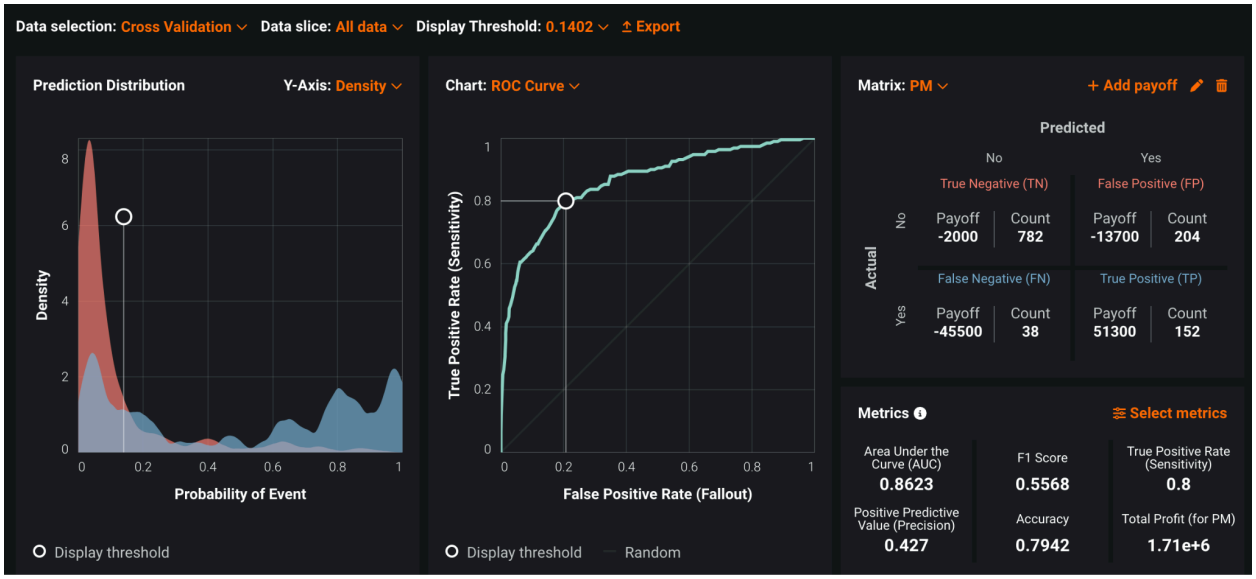

## *KNN Classifier*



## *SVM Classifier*




## *Neural Network*

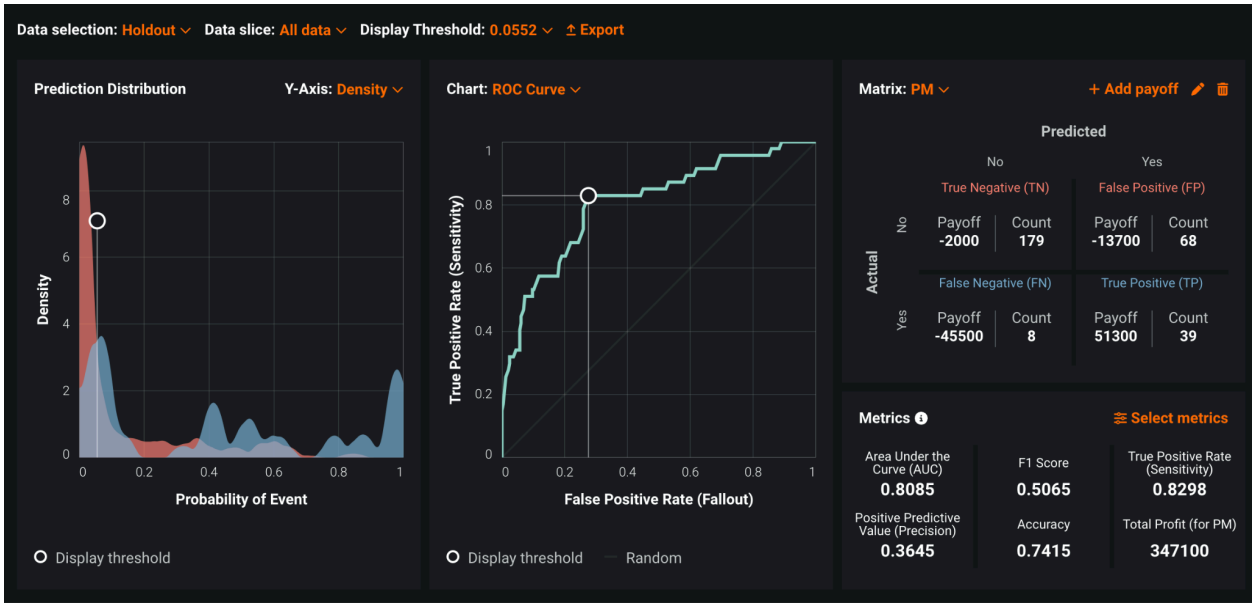


| | Cross Validation Metrics at Maximum Payoff | | | | | |
|---|---|---|---|---|---|---|
| | *Logistic Regression* | *Random Forest* | *Boosted Trees* | *KNN* | *SVM* | *Neural Network* |
| *Recall* | 0.7947 | 0.8 | 0.7632 | 0.7895 | 0.8211 | 0.8 |
| *Precision* | 0.439 | 0.3121 | 0.4215 | 0.2907 | 0.3861 | 0.427 |
| *F1* | 0.5655 | 0.449 | 0.5431 | 0.4249 | 0.5253 | 0.5568 |
| *Accuracy* | 0.8027 | 0.6828 | 0.7925 | 0.6548 | 0.7602 | 0.7942 |
| *ROC AUC* | 0.8611 | 0.7979 | 0.8388 | 0.7836 | 0.8595 | 0.8623 |
| *Maximum Payoff* | $1,740,000 | $1,77,100 | $1,090,000 | -$3,79,200 | $1,580,000 | $1,710,000 |
| *Threshold* | 0.1567 | 0.1303 | 0.1782 | 0.14 | 0.1226 | 0.1402 |

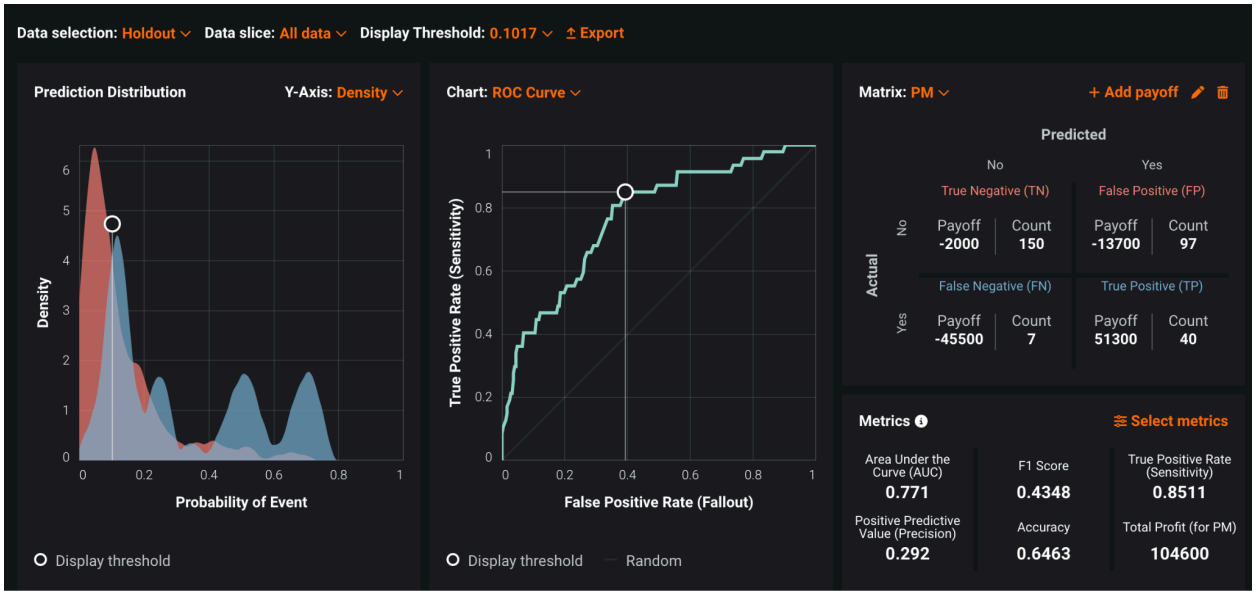
For cross validation, the payoff of Random Forest is the most. But, considering all the other metrics like precision and ROC AUC, logistic regression performs better. The payoff metric's difference for logistic regression and random forest is significant. There are some negative values which are seen because of imbalance in the payoff matrix.

**Holdout comparison of all models:**
All the values are captured by keeping the maximum payoff.

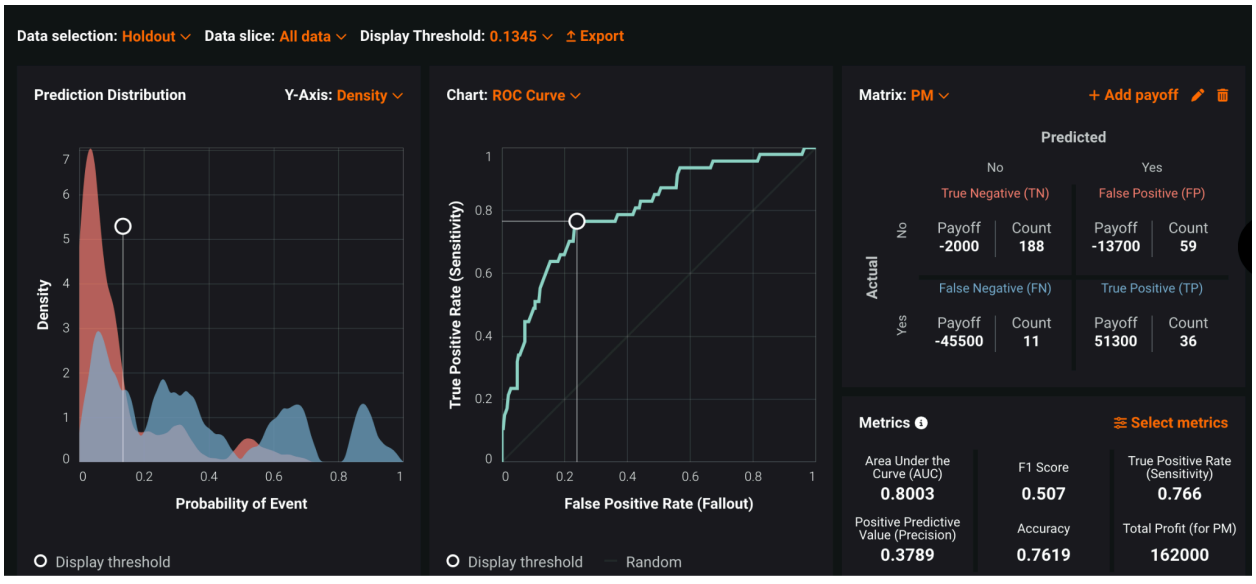
*Logistic Regression*



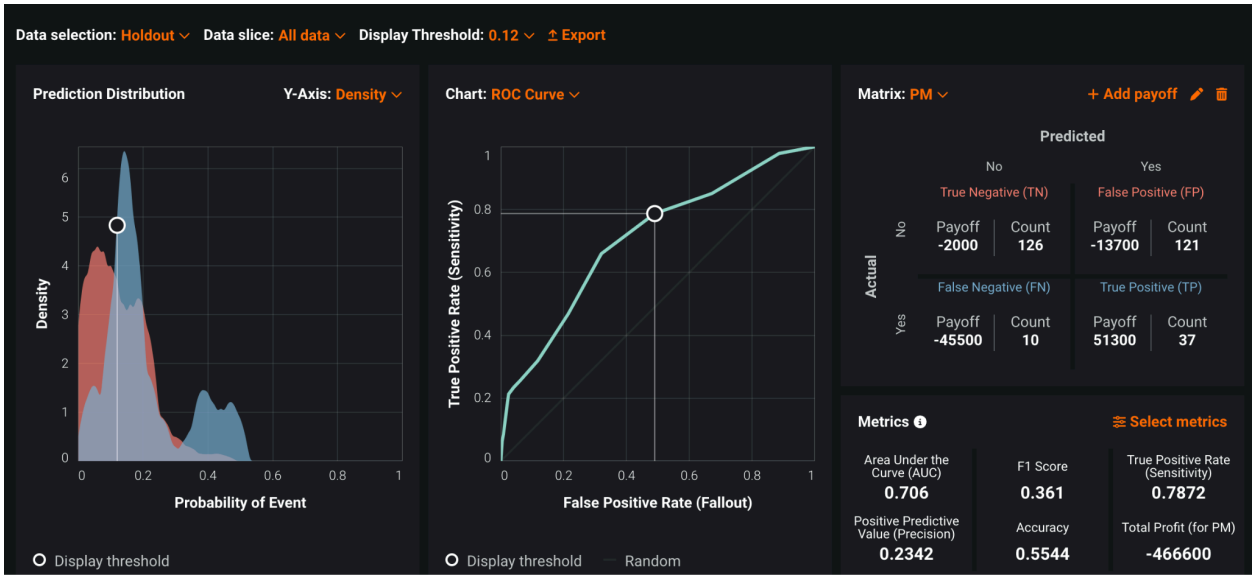
## *Random Forest Classifier*



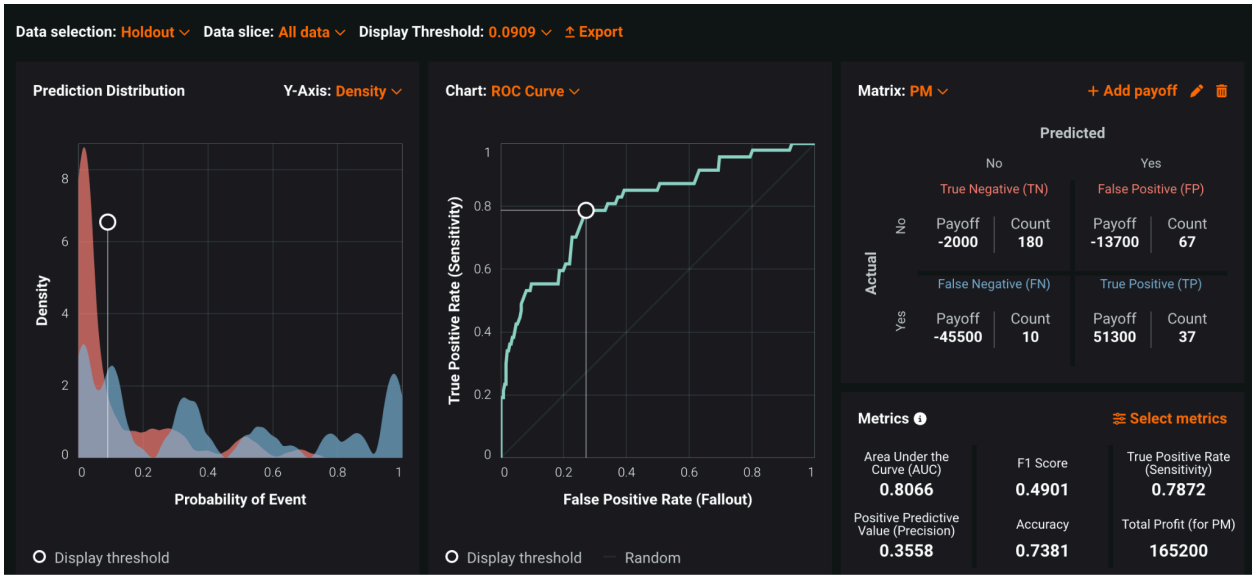

## *Boosted Trees Classifier*



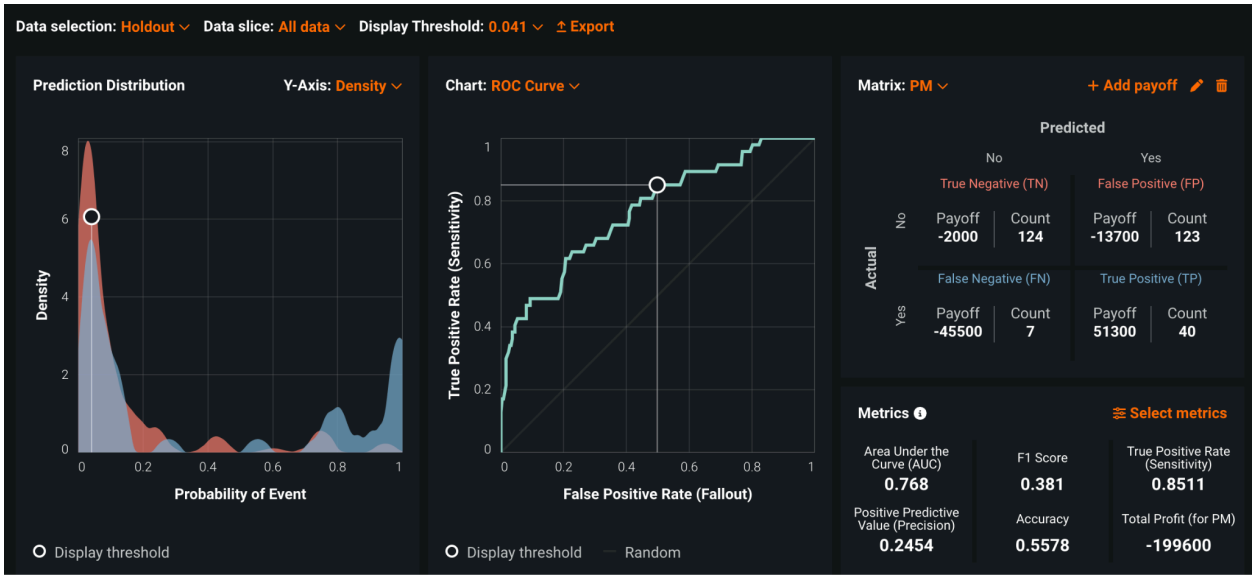
## *KNN Classifier*




## *SVM Classifier*



*Neural Network Classifier*




| | **Holdout Metrics at Maximum Payoff** | | | | | |
|---|---|---|---|---|---|---|
| | *Logistic Regression* | *Random Forest* | *Boosted Trees* | *KNN* | *SVM* | *Neural Networks* |
| *Recall* | 0.8298 | 0.8511 | 0.766 | 0.7872 | 0.7872 | 0.8511 |
| *Precision* | 0.3645 | 0.292 | 0.3789 | 0.2342 | 0.3558 | 0.2454 |
| *F1* | 0.5065 | 0.4348 | 0.507 | 0.361 | 0.4901 | 0.381 |
| *Accuracy* | 0.7415 | 0.6463 | 0.7619 | 0.5544 | 0.8066 | 0.5578 |
| *ROC AUC* | 0.8085 | 0.771 | 0.8003 | 0.706 | 0.7381 | 0.768 |
| *Maximum Payoff* | $3,47,100 | $1,04,600 | $1,62,000 | -$4,66,600 | $1,65,200 | -$1,99,600 |
| *Threshold* | 0.0552 | 0.1017 | 0.1345 | 0.12 | 0.0909 | 0.041 |

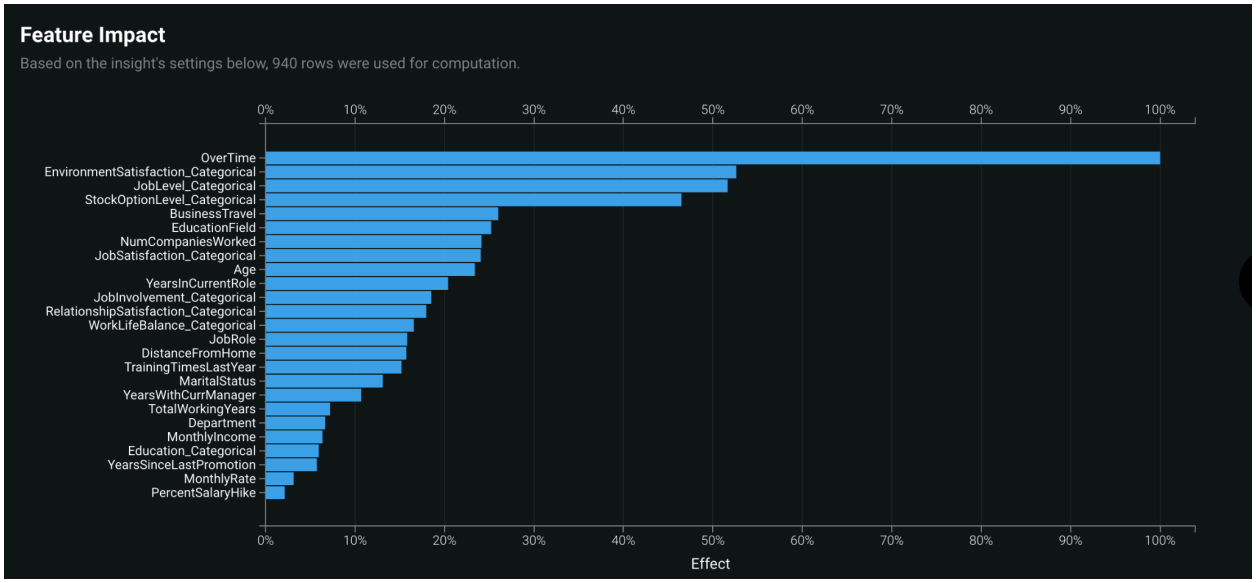
For holdout, logistic regression has the most payoff amount. All the metrics of logistic regression perform better than others.

There are some negative values which are seen because of imbalance in the payoff matrix.

**Q4. Top 4 predictors of customer churn.**

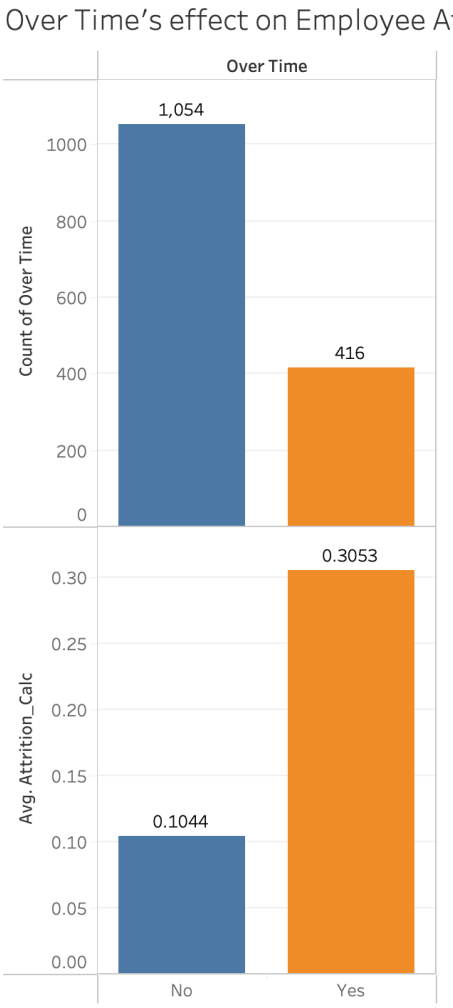
The best performing model was logistic regression and according to the same model, the features that impact attrition the most are selected.

The top 5 here are OverTime, EnvironmentalSatisfaction_Categorical, JobLevel_Categorical, StockOptionLevel_Categorical and Business Travel.



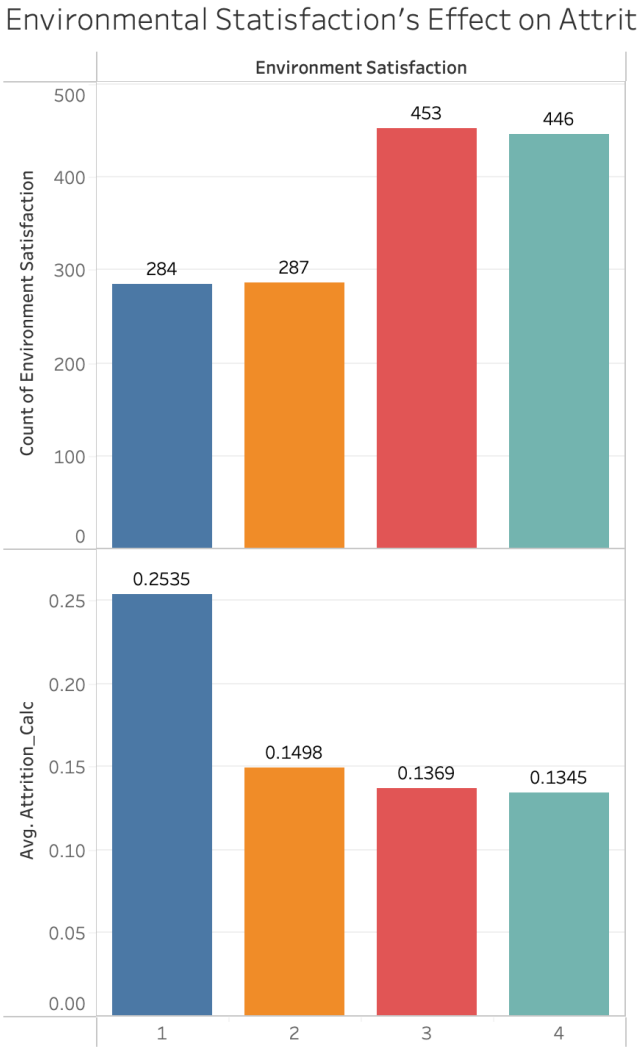

Feature impact list

**1st Feature -** OverTime



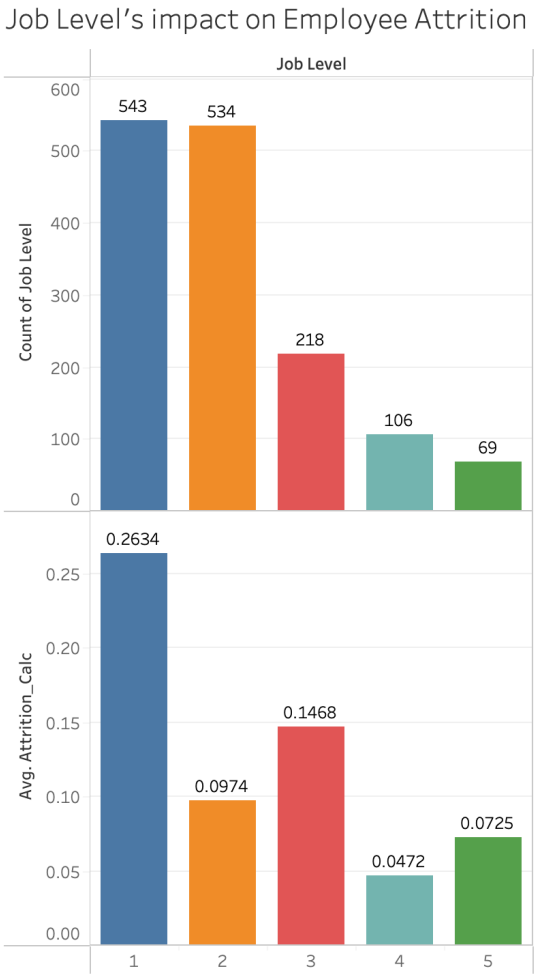

The people that work overtime are more likely to become attrition than the employees that don't work overtime.

**2nd Feature -** EnvironmentalSatisfaction_Categorical



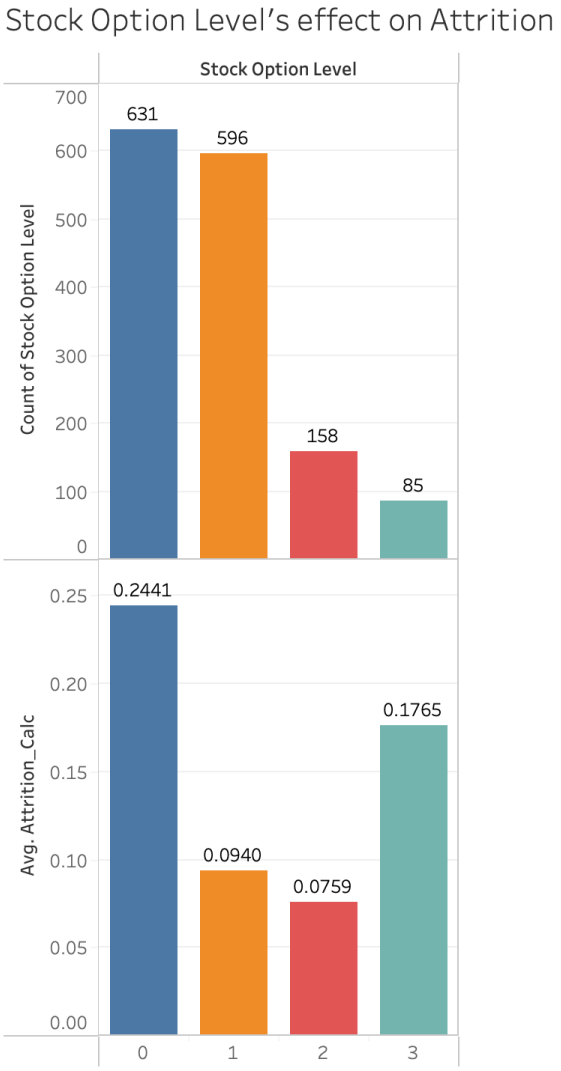

The attrition rate of the employees decreases with increase in environmental satisfaction, meaning the better the score for the environmental satisfaction, the less likely the person is to attrition.

**3rd Feature -** JobLevel_Categorical



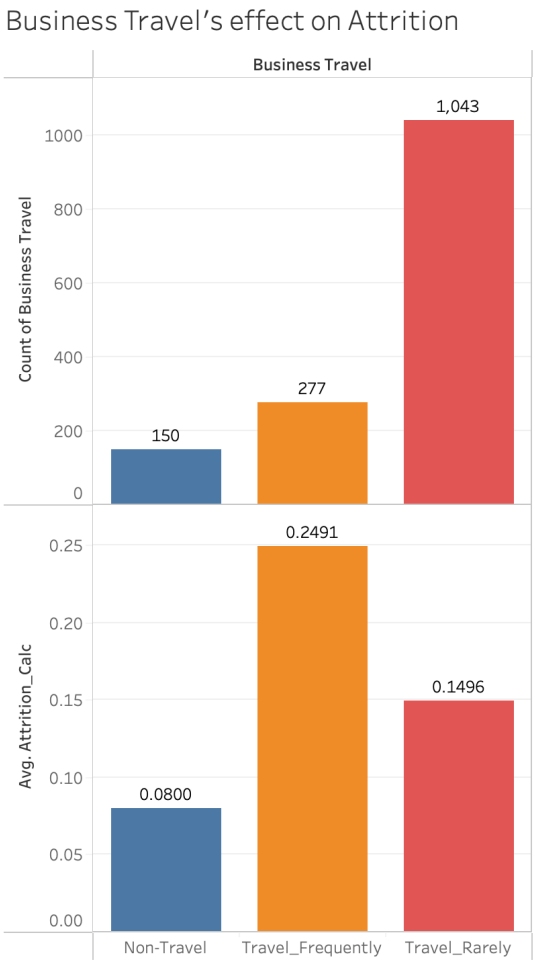

With increase in job level, the attrition rate decreases, but at level 3, some increase in attrition can be seen.

**4th Feature -** StockOptionLevel_Categorical




The employees that have the least stock option levels are the most likely to attrite.

**5th Feature -** Business Travel




The attrition rate of the employees that travel frequently are 10% more likely to attrition than the employees that travel rarely and 3 times more than non travelers.

Even though the number of employees that travel rarely are almost 3.7 times more than the travelers, their attrition rate is lower.

**Q5. Actionable insights on top 5 predictors of customer churn.**

**1st Feature - OverTime**
The employees working overtime can be provided with a better compensation for the extra work time.

**2nd Feature - EnvironmentalSatisfaction_Categorical**
Research should be done on which factors are affecting the environmental satisfaction for the employees and then, further steps should be taken. From this data, we can't point towards any particular thing that causes less environmental satisfaction hence there are no actionable insights other than doing further research.

**3rd Feature - JobLevel_Categorical**
Job levels are the positions that each employee holds in an organization, and here, there isn't any action that can be taken.

**4th Feature - StockOptionLevel_Categorical**
Employees that are provided with less stock options should be given more options so that they feel valued and continue to stay.

**5th Feature - Business Travel**
Employees that travel more are likely to attrition more as they may move to a different location, or otherwise. Having packages that make the employees stay with bonuses and retention strategies could be tried but there is no guarantee that they would stay.