Attrition Class Prediction Using Machine Learning Algorithm and Feature Engineering to Enhance Model Accuracy

UCHECHUKWU FREDRICK OKONKWO

05TH AUGUST 2022

9 WEEKS DURATION

OVERVIEW

- PROBLEM STATEMENT
- OBJECTIVE
- RELATED STUDIES
- DATA SET, EDA & VISUALIZATION
- METHODOLOGY & EXPERIMENTAL SETUP
- RESEARCH RESULTS
- CHALLENGES AND IN PROGRESS

PROBLEM STATEMENT

Human resource analytics is a multi-disciplinary field that incorporates various methodologies to improve the quality of people-related decisions which in turn enhances organizational performance.

It is essential to gain this information because recruiting employees is often arduous and costlier than retaining old ones. Therefore, companies make use of this information to see how they can prevent the employee from leaving by improving the employee retention.

If the attrition rate is high, that signifies a red flag which calls for proper scrutiny to identify the major underlying factors that contributes more to that phenomenon and those that don’t.

ATTRITION PREDICTION ?

Attrition prediction is essentially predicting the employees that are most likely to quit their job based on specific factors, for instance, age, performance rating, distance of work to home, salary increase, department, marital status, mode of work, education and so on

OBJECTIVE

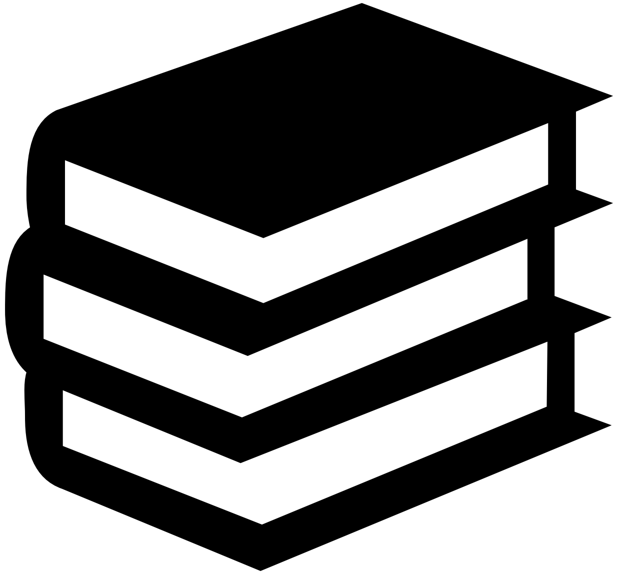
In this project, I combined machine learning techniques and advanced feature engineering techniques to predict if an employee would quit his/her job or not eventually and finally evaluate and select the best model based on the following model performance metric:

1) Accuracy
2) Precision
3) Recall
4) F1- score
5) ROC-AUC
6) Latency (computational duration)

ATTRITION PREDICTION

Questions to address:

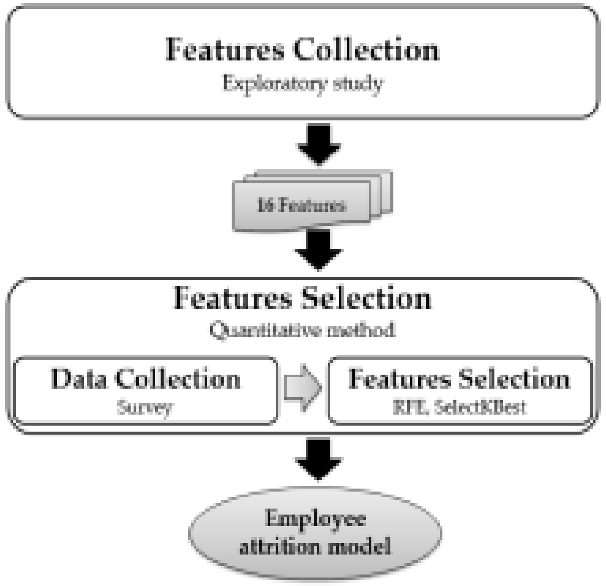
- **What factors or features have most influence on the attrition rate of employees and how can they be identified?**
- **Does the combination of ensemble learning, feature engineering, SMOTE and feature selection enhance a model's predictive performance or otherwise?**



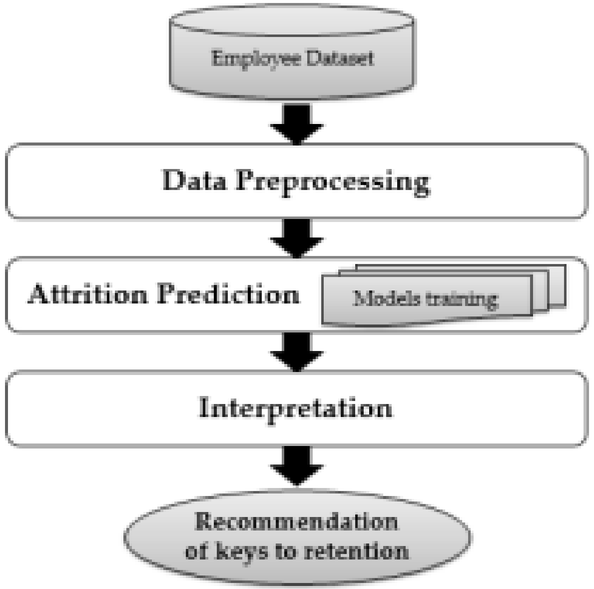
RELATED WORK

N.B. Yahia et al (2021), in this attrition prediction applied feature collection by exploratory study of related scientific research and qualitative feature selection, recursive feature elimination (wrapper method) and SelectKbest (filter method). Classifiers used are decision tree, SVM, logistic regression, random forest(ensemble learning), XGBoost, Voting Classifier and stacked ANN-based model.
VC gives the best result with accuracy of 93% followed by the RF classifier with 85.8% then XGB with 85.3%. All ensembling methods.
NB: no SMOTE

N.B. Yahia et al. feature selection mixed method flow chart



Architecture of the proposed approach



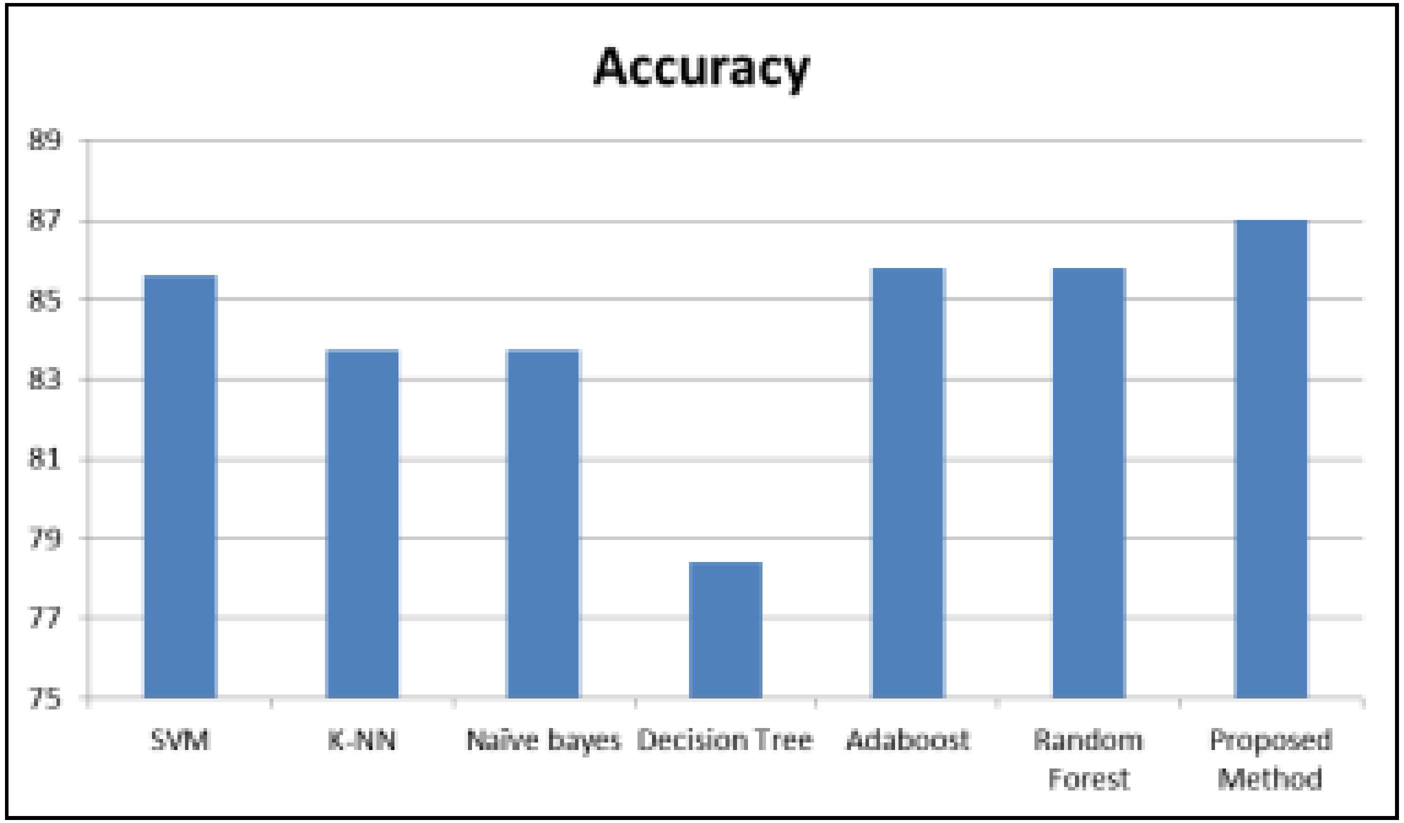
Models	DT	LR	SVM	DNN	RF	XGB	VC
			IBM HR Dataset				
[14]			0.74		0.71		
[15]		0.89			0.87	0.87	
[16]	0.82	0.87	0.86	0.86	0.86	0.86	
[17]		0.9			0.92		
[18]			0.85				
[19]		0.81	0.77		0.82		0.83
[22]	0.69	0.86	0.87				0.88
[23]	0.79	0.86			0.85	0.9	
[24]	0.85						
[25]	0.83						
Ours	0.77	0.83	0.85	0.8	0.858	0.853	**0.93**

Shawna Dutta et al(2020). proposed system implements the use of feed-forward neural network with 10-fold cross validation procedure under a single platform to predict attrition probability. This method was evaluated and compared with 6 classifiers such as SVM, K-Nearest neighbors, naïve bayes, decision tree, Adaboost and RF classifiers. The eural network method with 10 fold cross- validation achieved maximum performance of 87% compared to other classifiers.

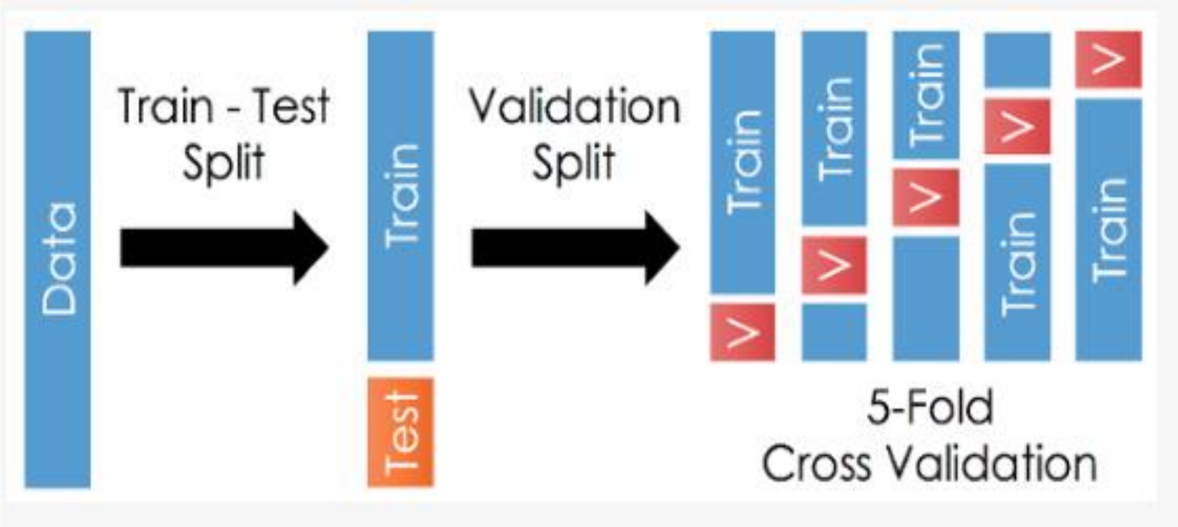
Performance comparison of all specified baseline classifiers

Performance Measure Metrics	SVM	K-NN	Naïve Bayes	Decision Tree	Adaboost	Random Forest
Accuracy	85.6%	83.74%	83.74%	78.4%	85.8%	85.8%
MSE	0.144	0.1626	0.16	0.216	0.14	0.142

Overall Performance of the classifiers with respect to Accuracy



K_fold cross validation. (F. Fallucchi et al 2020)

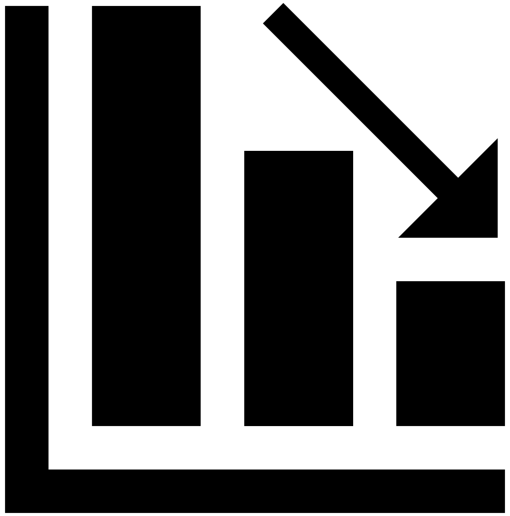


Gaussian Naïve Bayes confusion matrix. (F. Fallucchi et al 2020)

	Predicted 0	Predicted 1	
Real 0	313 70.98%	57 12.93%	370 84.59% 15.41%
Real 1	20 4.54%	51 11.56%	71 71.83% 28.17%
	333 93.99% 6.01%	108 47.22 % 52.78%	441 82.54% 17.46%

Evaluation metric results for all classifiers used. (F. Fallucchi et al 2020.)

	Accuracy Train	Accuracy Test	Precision	Recall	Specificity	F1 Score
Gaussian NB	0.782	0.825	0.386	0.541	0.845	0.446
Bernoulli NB	0.831	0.845	0.459	0.331	0.927	0.379
Logistic Regression	0.865	0.875	0.663	0.337	0.962	0.445
K Nearest Neighbour	0.842	0.852	0.551	0.090	0.994	0.150
Decision Tree	0.792	0.823	0.356	0.361	0.910	0.351
Random Forest	0.850	0.861	0.658	0.132	0.991	0.194
SVC	0.851	0.859	0.808	0.096	0.994	0.166
Linear SVC	0.858	0.879	0.665	0.247	0.978	0.358



DATA SET, EDA & VISUALIZATION

```
Age                          int64
Attrition                   object
BusinessTravel              object
Department                  object
DistanceFromHome             int64
Gender                      object
JobInvolvement               int64
JobLevel                     int64
JobRole                     object
JobSatisfaction              int64
MaritalStatus               object
MonthlyIncome                int64
NumCompaniesWorked           int64
OverTime                    object
PercentSalaryHike            int64
PerformanceRating            int64
StockOptionLevel             int64
TotalWorkingYears            int64
TrainingTimesLastYear        int64
YearsAtCompany               int64
YearsSinceLastPromotion      int64
YearsWithCurrManager         int64
Higher_Education            object
Status_of_leaving           object
Mode_of_work                object
Leaves                       int64
Absenteeism                  int64
Work_accident               object
Source_of_Hire              object
Job_mode                    object
```

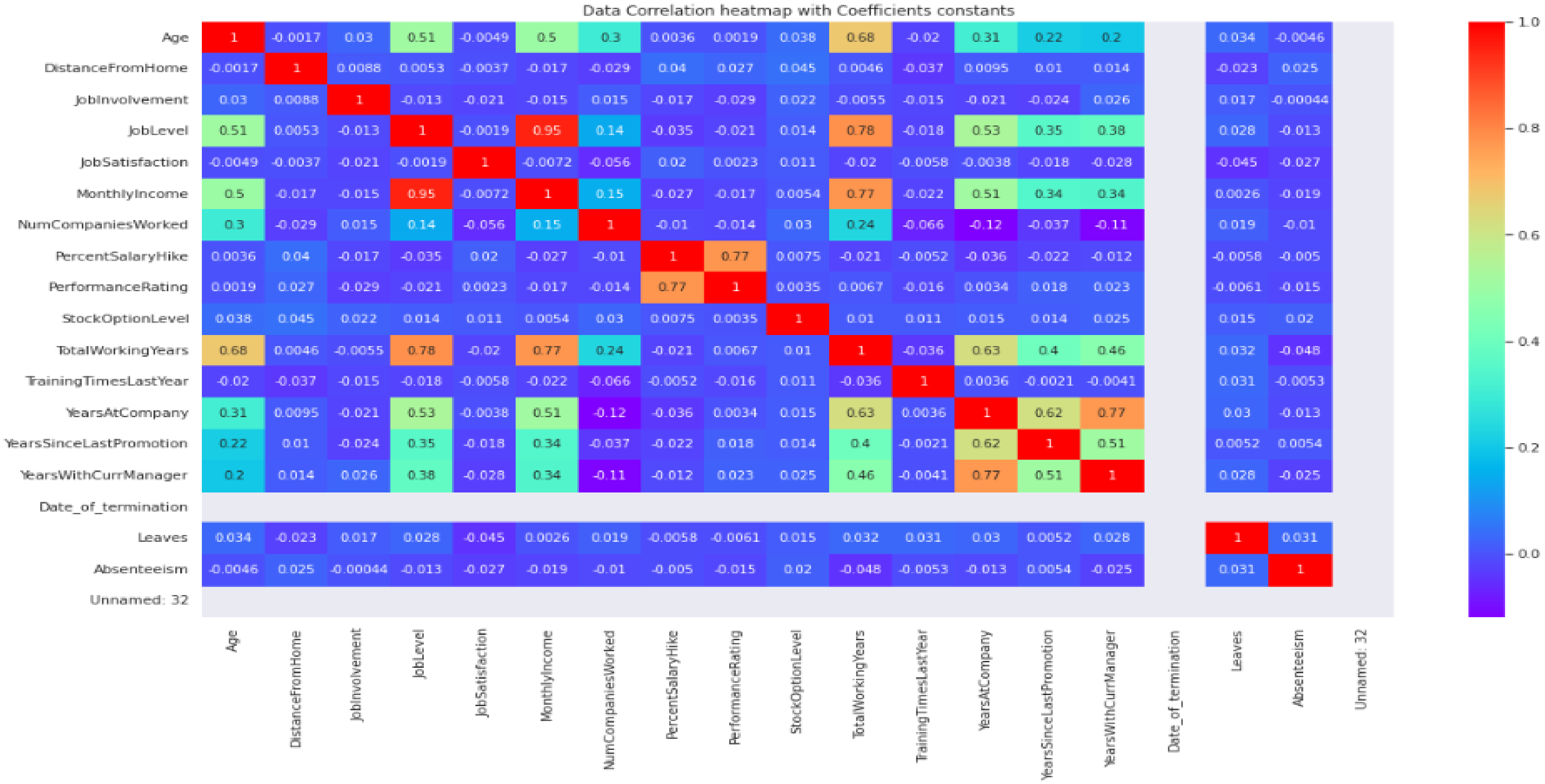
HR Attrition data based on IBM attrition.

Dataset was gotten from Kaggle which is a public online data repository.

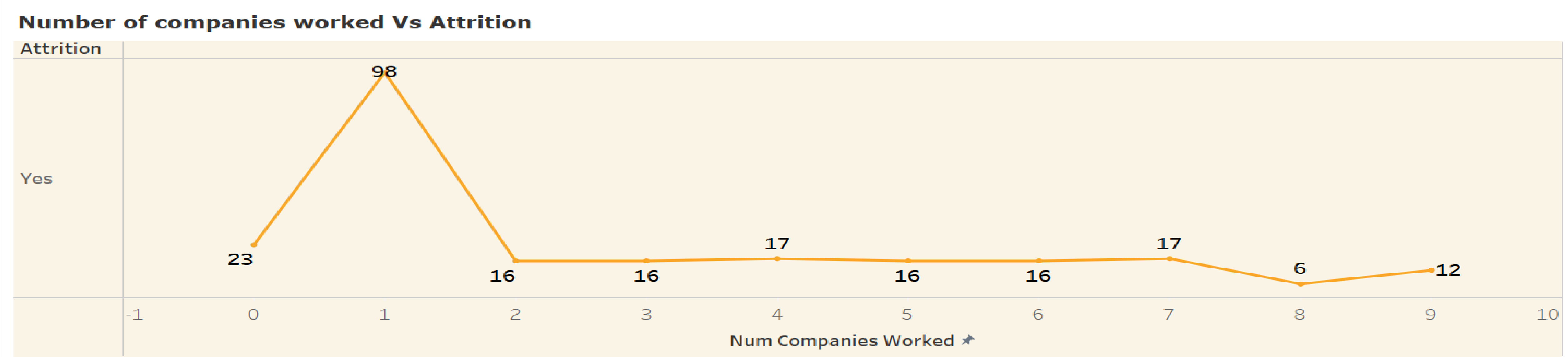
It consist of 32 columns with 1470 instances.

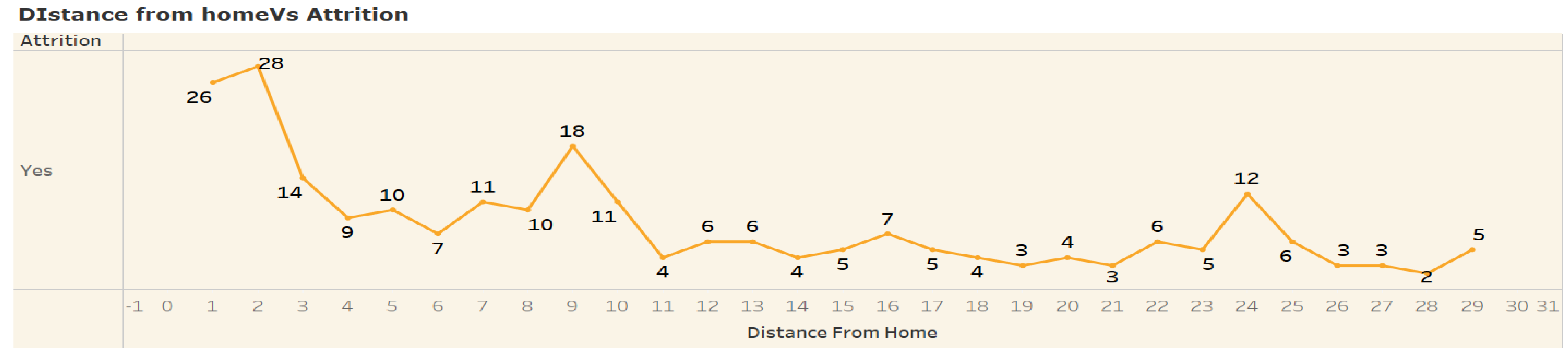
17 columns are numerical variable, and 13 columns are categorical variables, while 2 are empty columns, respectively.

Pearson's Correlation coefficient heatmap

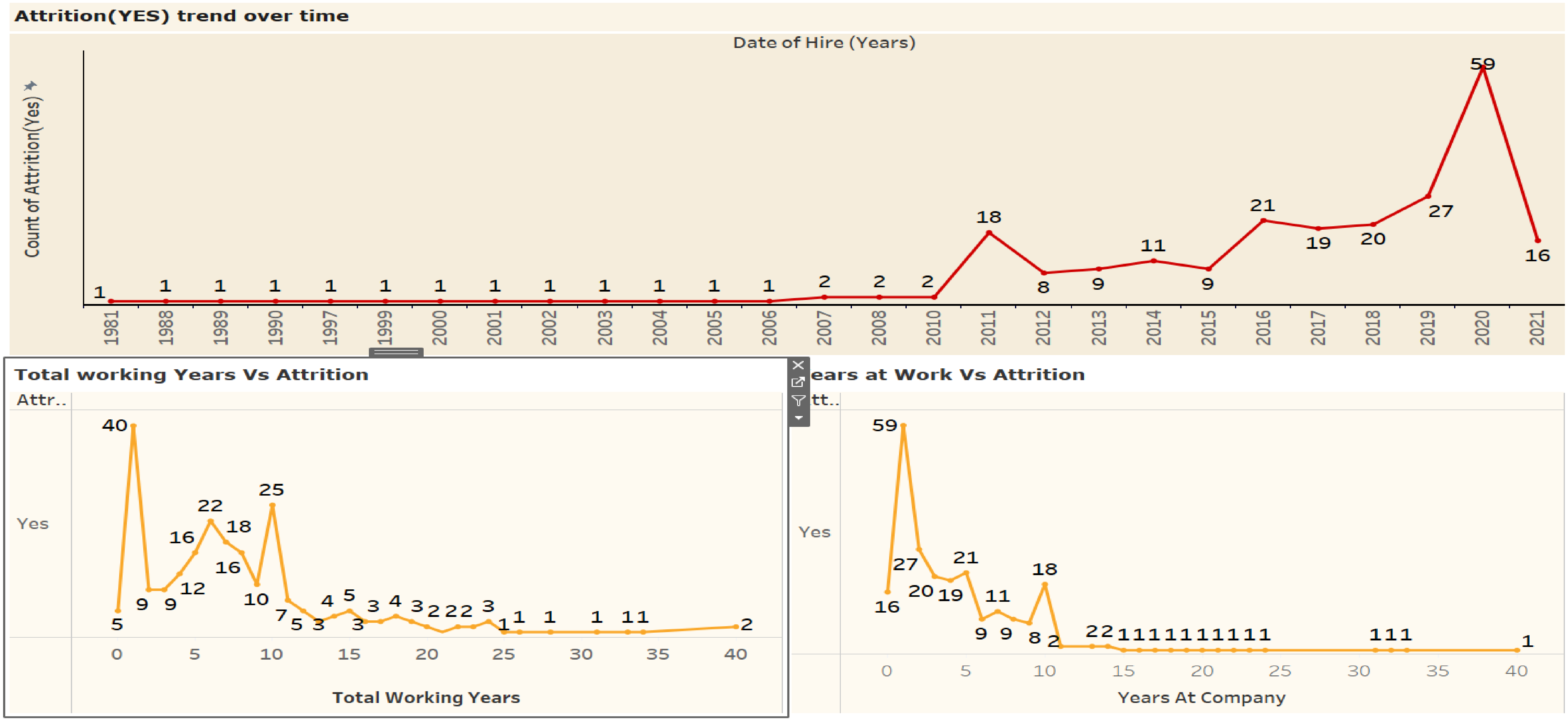


EXPLORATORY DATA ANALYSIS VISUALIZATIONS (TABLEAU) DASHBOARD 1

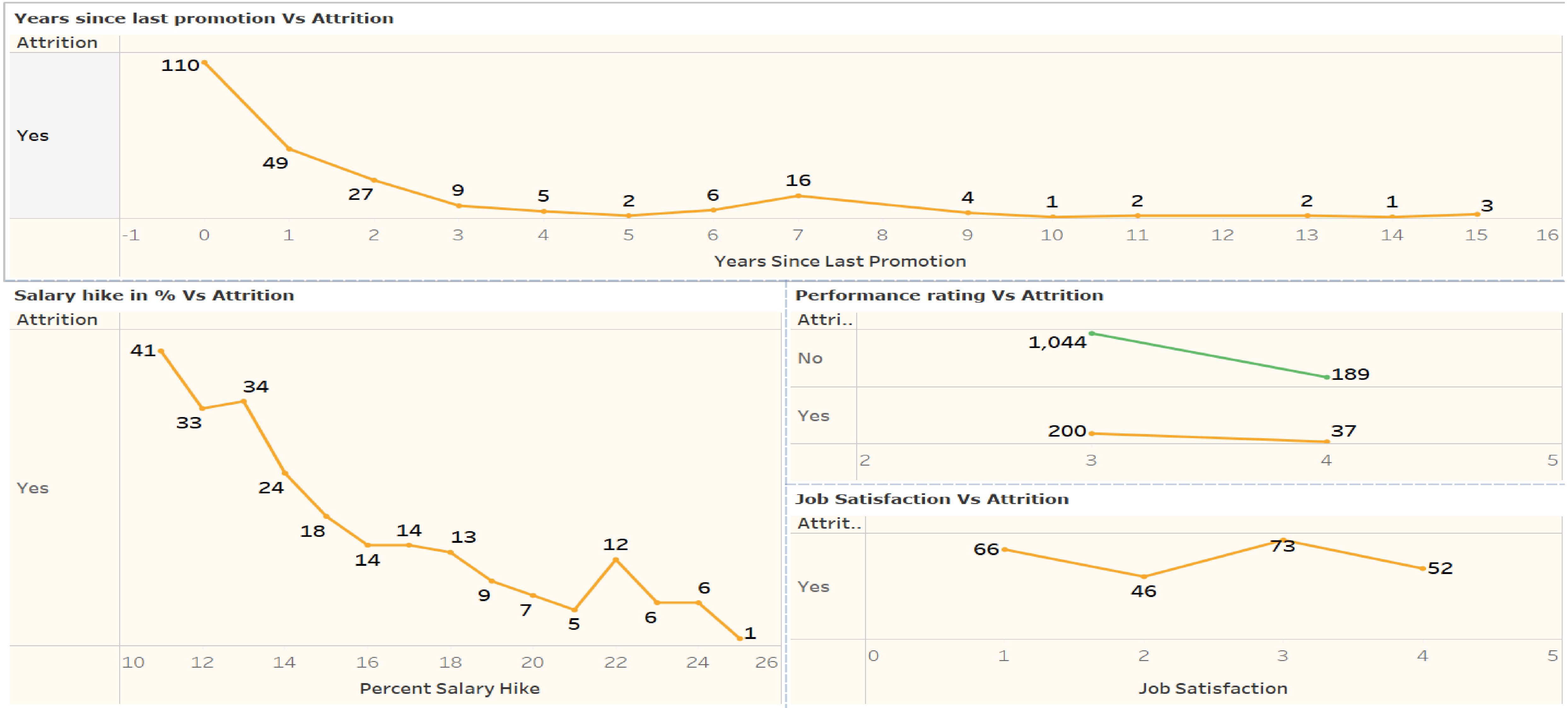




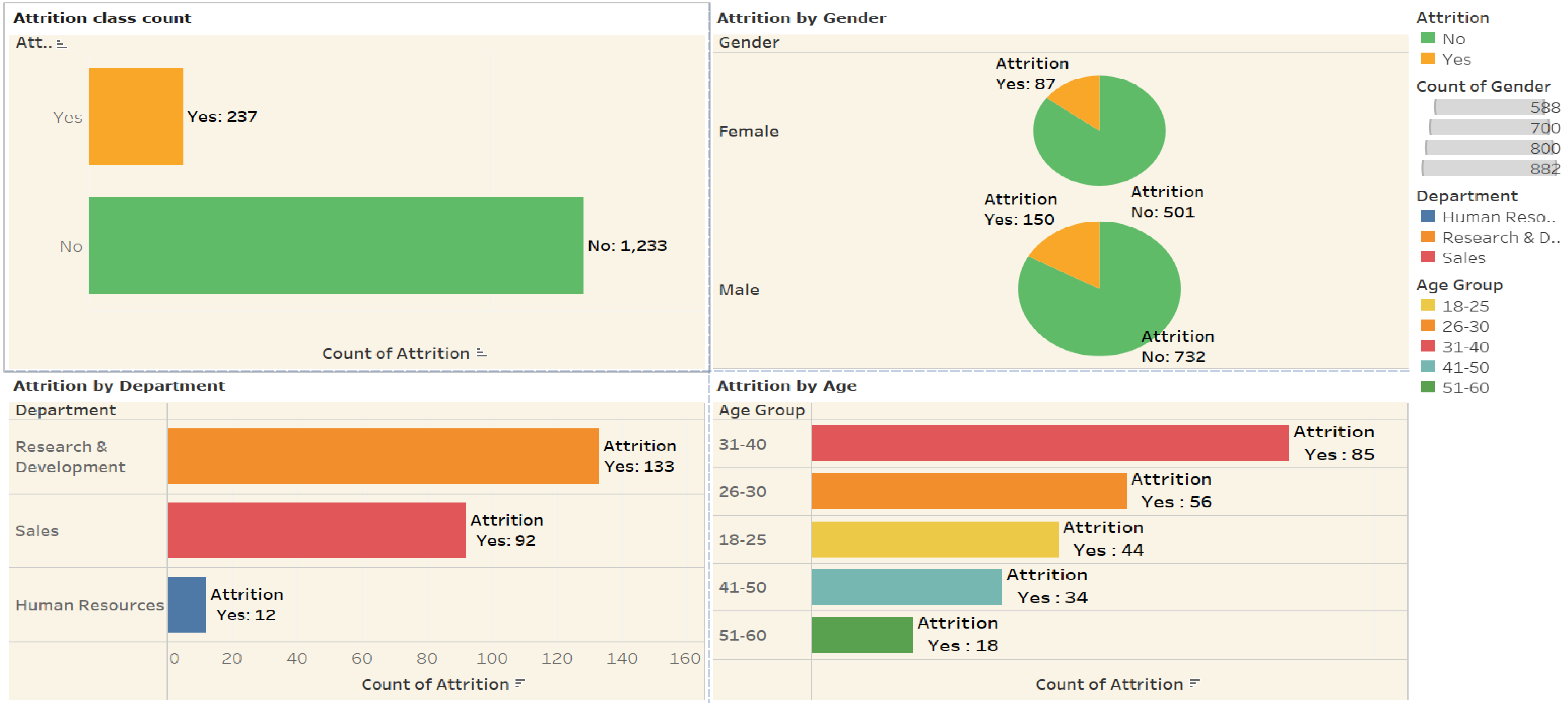
EXPLORATORY DATA ANALYSIS VISUALIZATIONS (TABLEAU) DASHBOARD 2



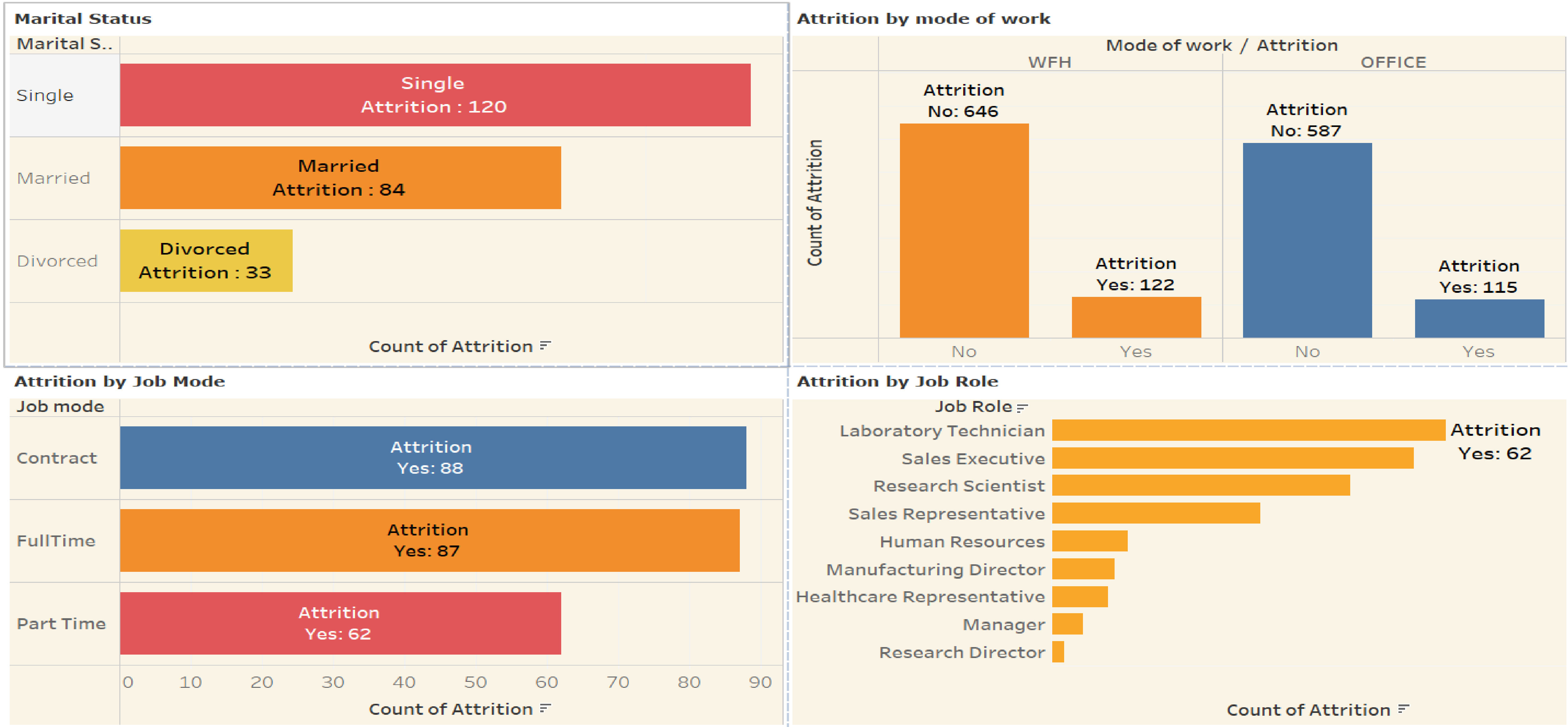
EXPLORATORY DATA ANALYSIS VISUALIZATIONS (TABLEAU) DASHBOARD 3



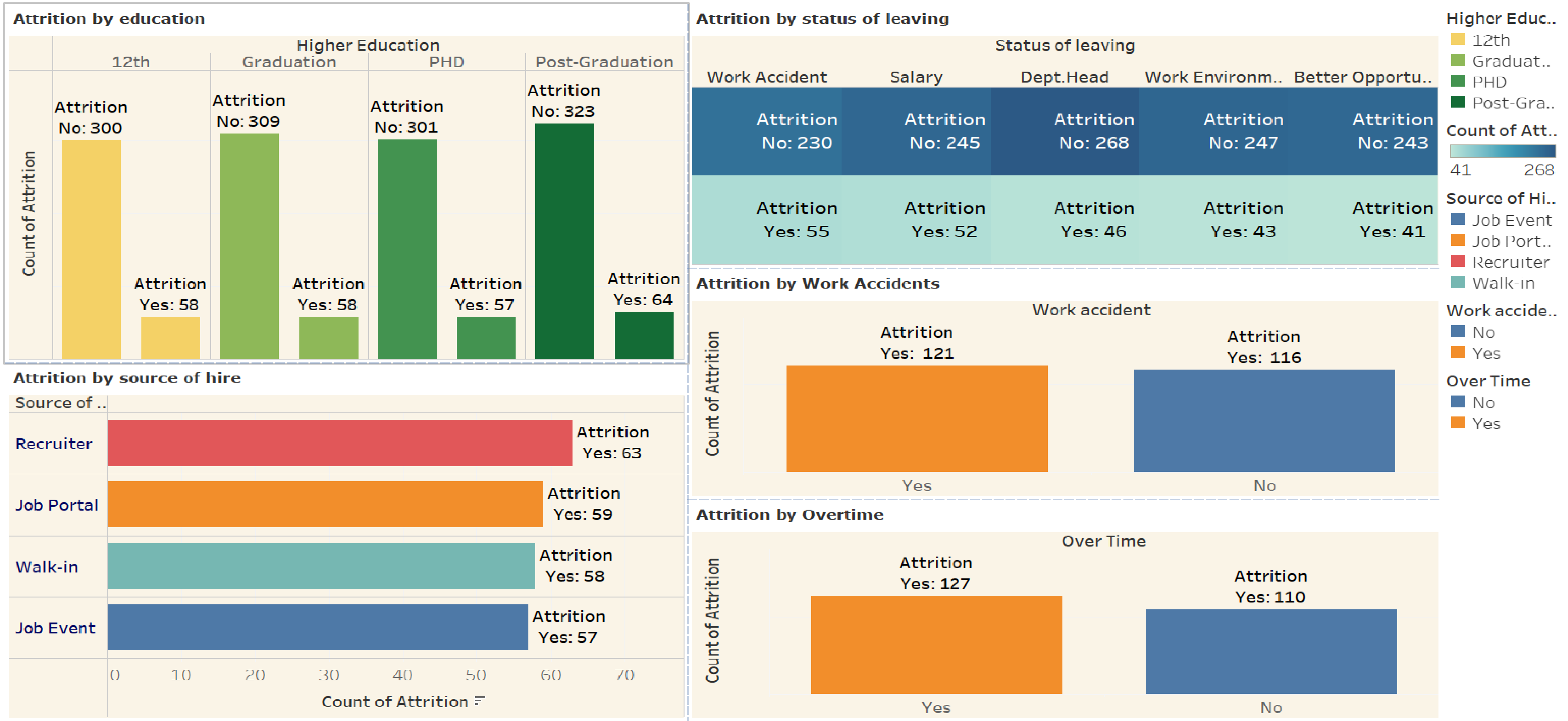
EXPLORATORY DATA ANALYSIS VISUALIZATIONS (TABLEAU) DASHBOARD 4

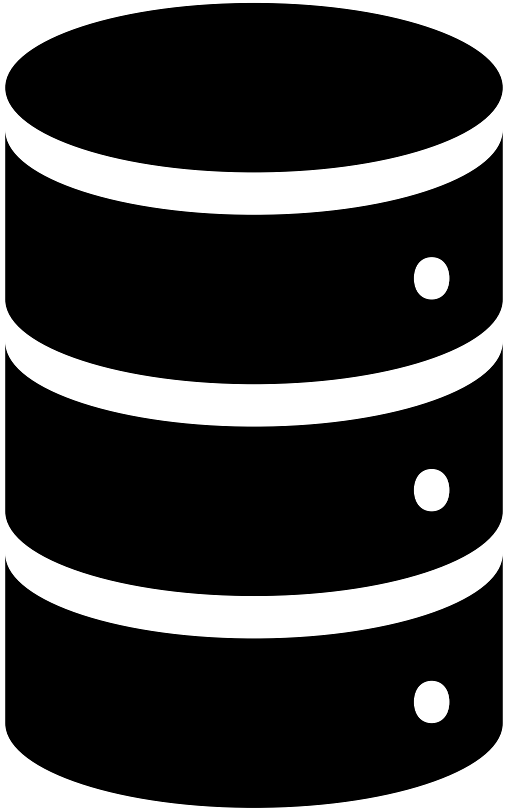


EXPLORATORY DATA ANALYSIS VISUALIZATIONS (TABLEAU) DASHBOARD 5



EXPLORATORY DATA ANALYSIS VISUALIZATIONS (TABLEAU) DASHBOARD 6





METHODOLOGY & EXPERIMENTAL SETUP

Software/Programming language: Python on Google Collaboratory

Machine Learning algorithms: Random forest classifier, support vector machine, Logistic regression.

Feature engineering: 6 sets of features

Preprocessing: data cleaning, feature selection, SMOTE (synthetic minority oversampling technique), box-cox transformation, encoding, standard scaling.

Data splitting: Train 60%, Validation 20%, test 20%

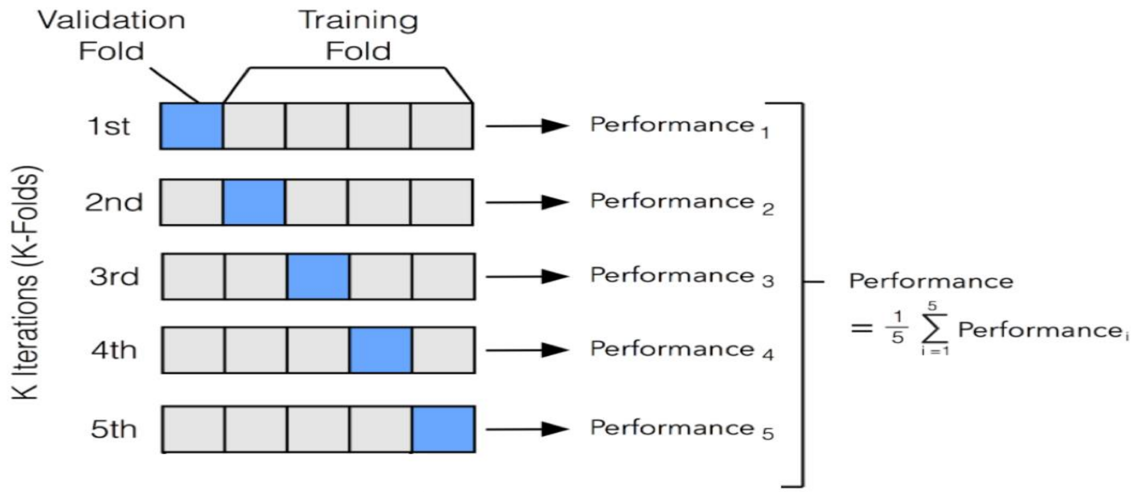
Model evaluation

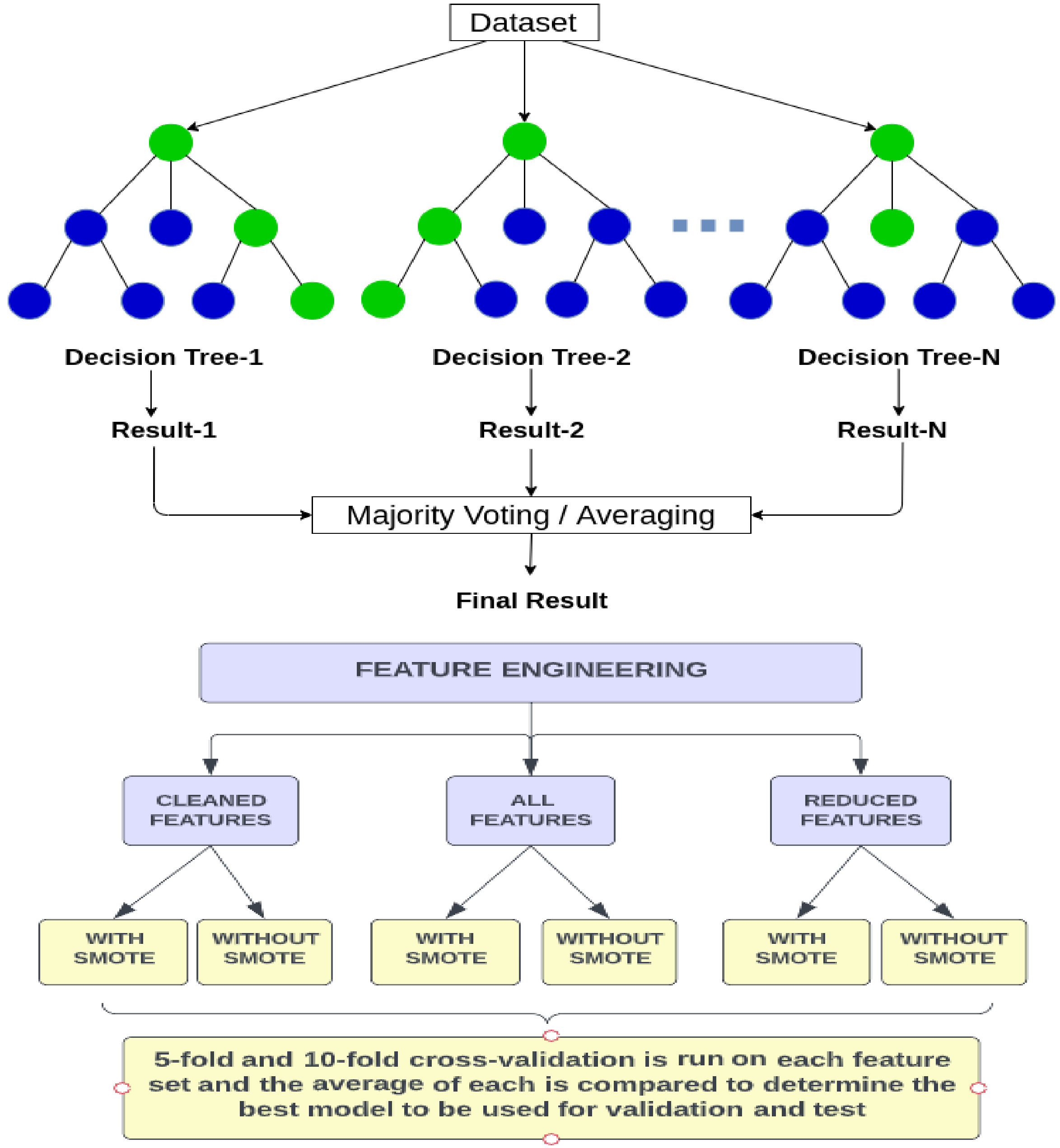
RANDOM FOREST CLASSIFIER (ensemble learning)

Max_depth: 2,4,8,16,32,none

Number of estimators: 2^i, where *i*=3-10

- **5-fold and 10- fold cross-validation was run on the feature set and the best model was selected.**
- **Evaluate the models on the validation set and pick the best one**
- **Evaluate the best model on the test set to gauge the model's ability to generalize to unseen data and to confirm its consistency.**





PREPROCESSING

Feature selection for categorical features:

Chi squared test (filter method)
We calculate Chi-square between each feature & the target & select the desired number of features with best Chi-square scores or the lowest p-values.

Categorical columns to check for predictive power
Attrition, Business Travel, Department, Gender, Job Role, Marital Status, Over Time, Higher Education, Status of leaving, Mode of work, Work accident, Source of Hire, Job mode

The Formula for Chi Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

c = degrees of freedom
O = observed value(s)
E = expected value(s)

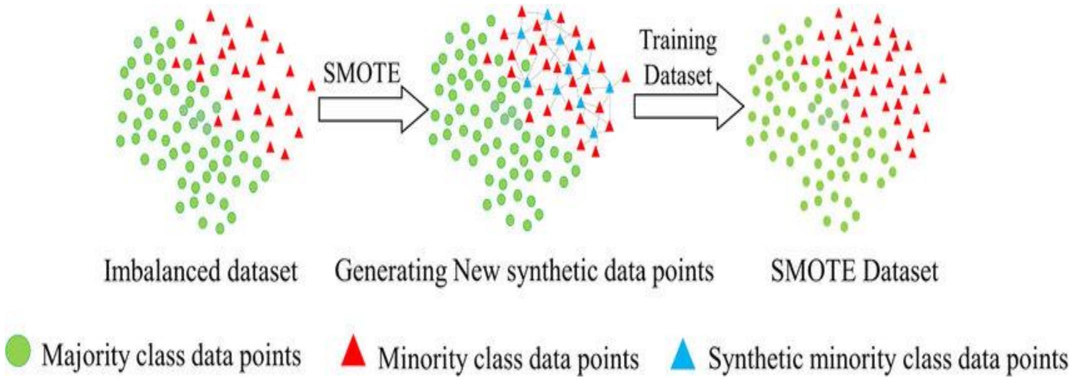
- Data cleaning:
Dropping empty columns.
Renaming columns.

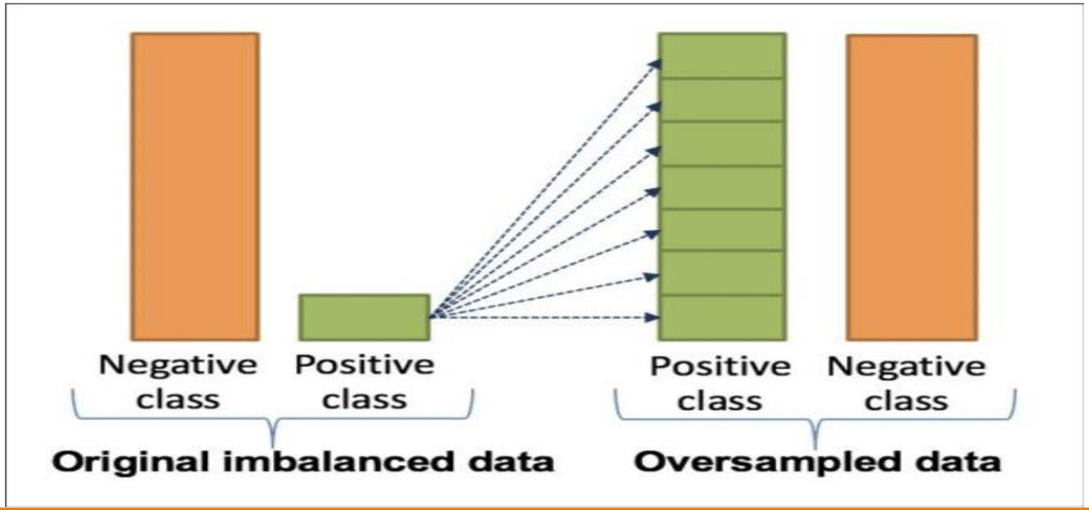
- Categorical data Encoding.

- Standard scaling

Synthetic Minority Oversampling Technique (SMOTE):
This techniques is employed to tackle data imbalance.

```
Before SMOTE: Counter({0: 735, 1: 147})
After SMOTE: Counter({0: 735, 1: 735})
```





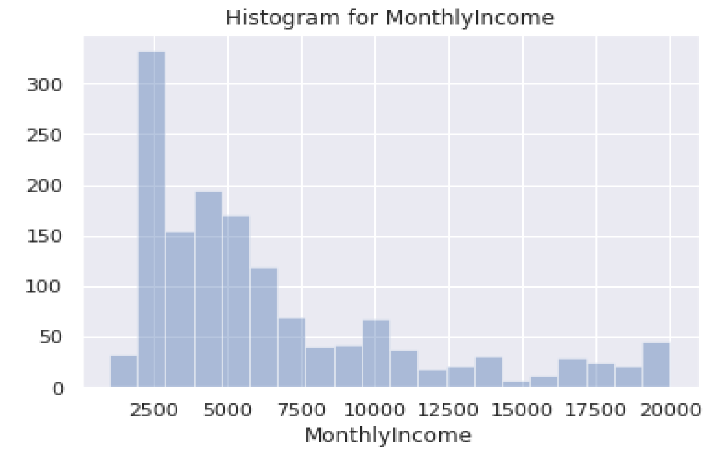
Data transformation:

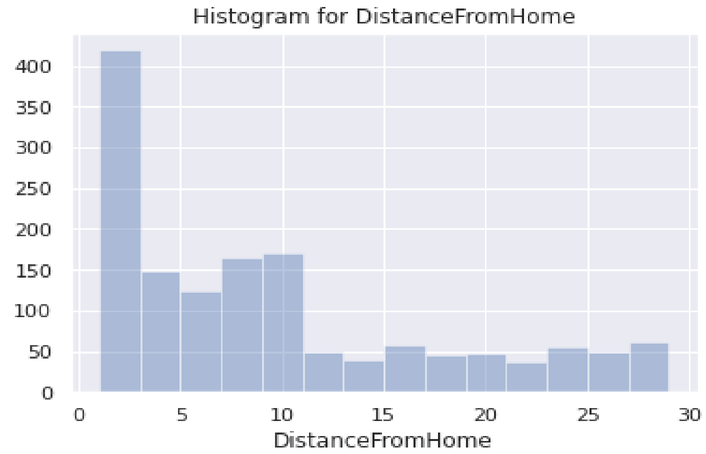
Box Cox transformation to convert skewed continuous features to a normally distributed one.

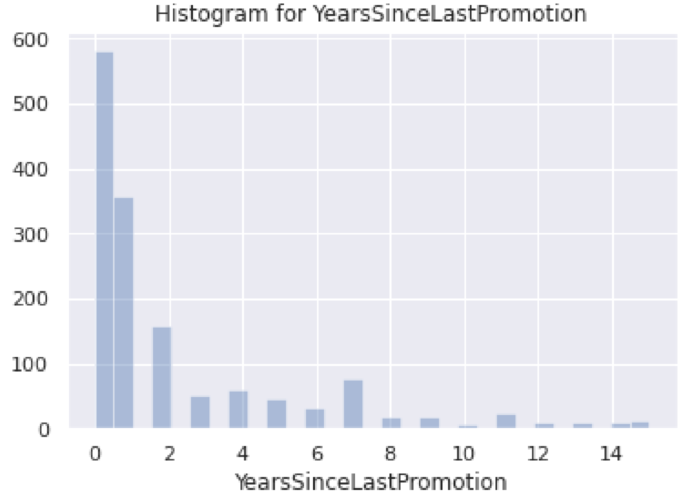
Columns to be transformed: Monthly Income, TotalWorkingYears, YearsAtCompany,

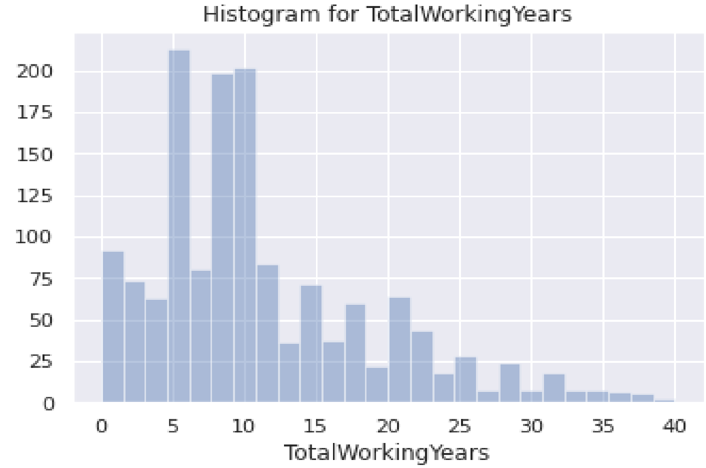
$$
y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(y_i) & \text{if } \lambda = 0, \end{cases}
$$

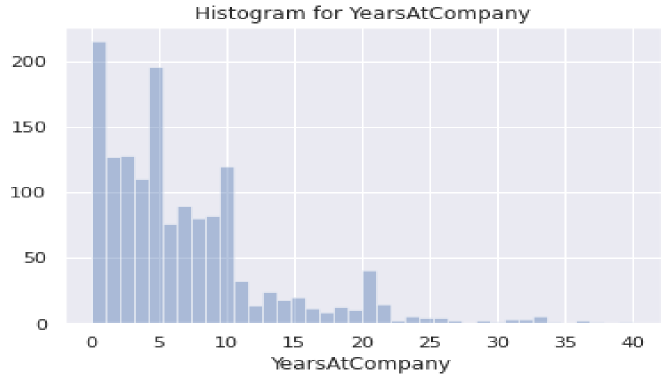
λ	Transformed Data
-2	y^{-2}
-1	y^{-1}
-0.5	$1/\sqrt{y}$
0	$\ln(y)$
0.5	$\sqrt{y}$
1	y
2	y^2

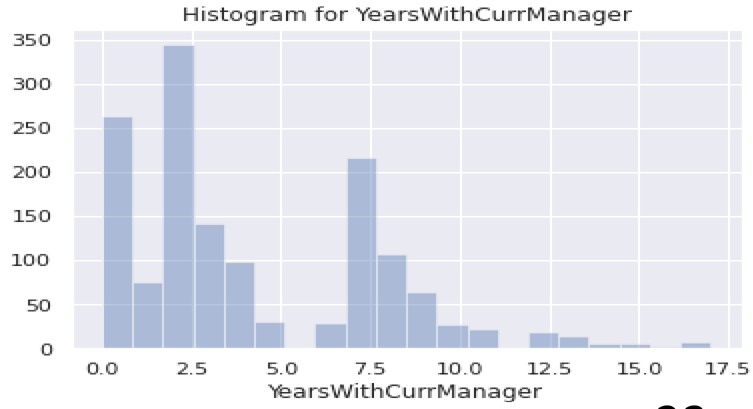




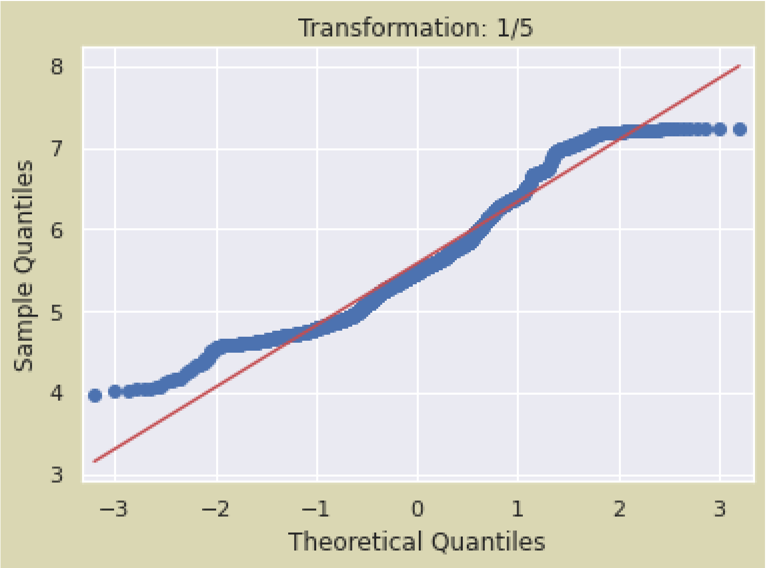


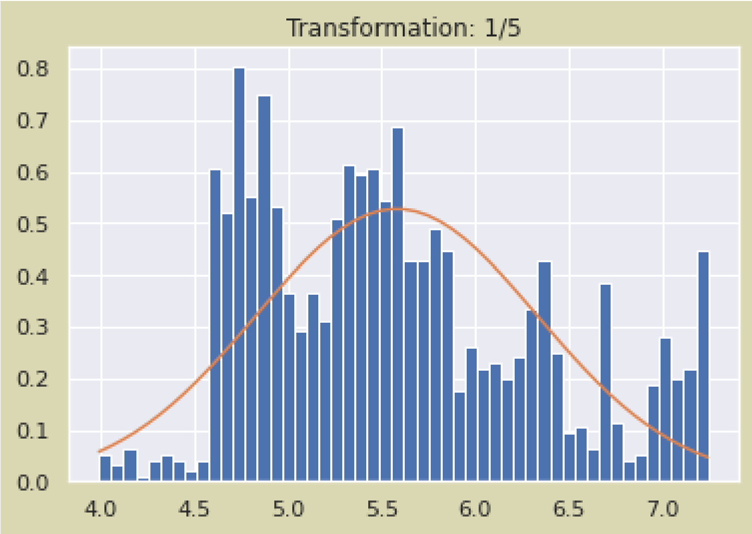




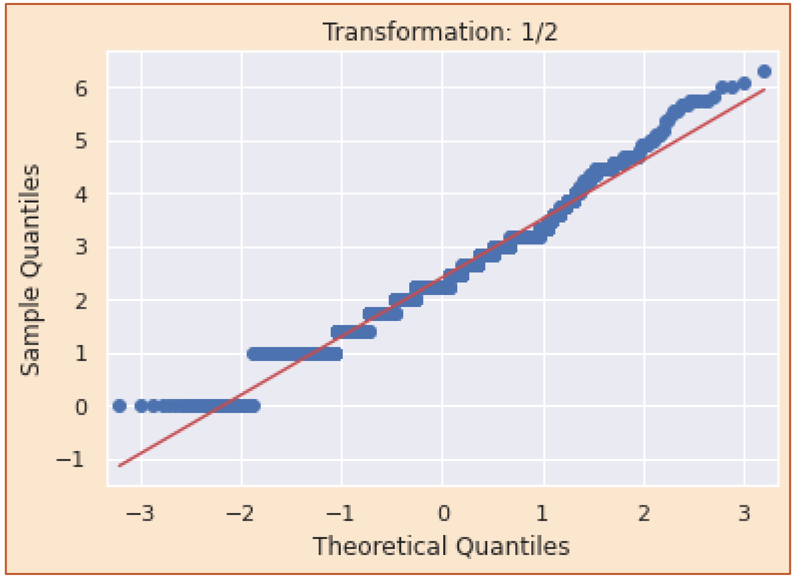


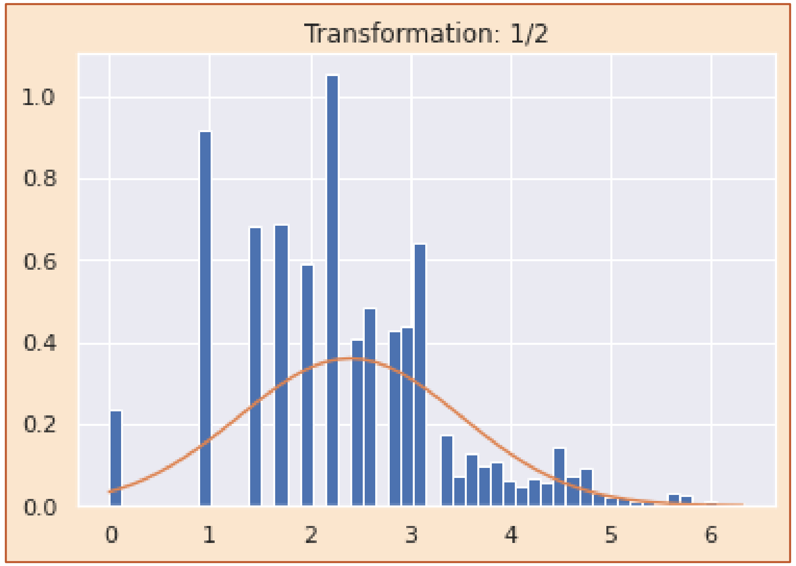
MONTHLY INCOME



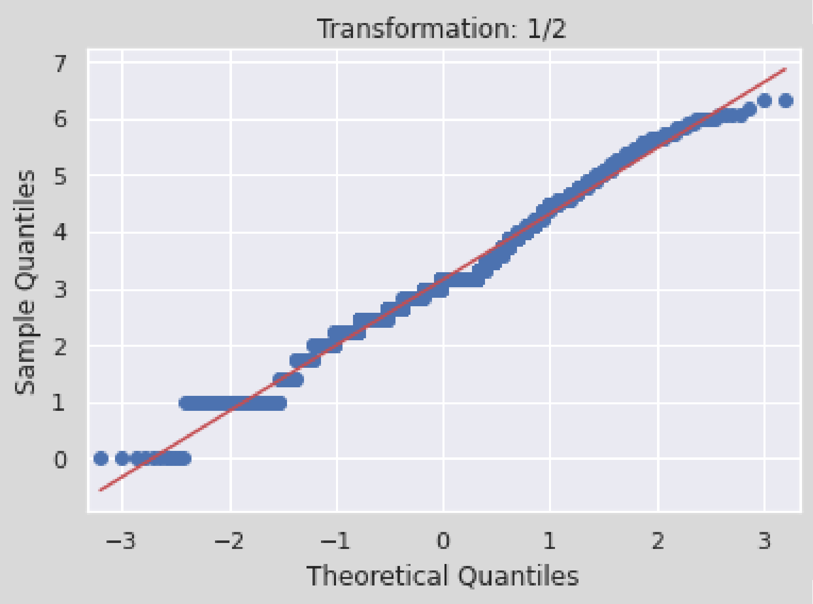


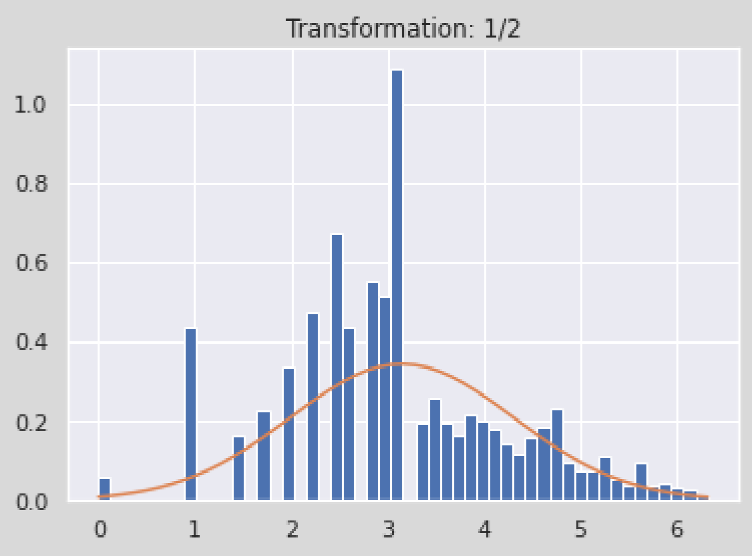
YEARSATCOMPANY





TOTALWORKINGYEARS





RESEARCH RESULTS

5-fold and 10- fold cross-validation was run on train 60% dataset for all 6 feature sets respectively.

RANDOM FOREST CLASSIFIER MODEL RESULTS 5-fold CV

Feature set	Max	Average	Min
SMOTE CLEAN	0.912	0.857	0.792
SMOTE ALL	0.916	0.854	0.79
SMOTE REDUCED	0.904	0.83	0.76
CLEAN	0.85	0.857	0.83
ALL	0.853	0.847	0.824
REDUCED	0.846	0.837	0.823

RANDOM FOREST CLASSIFIER MODEL RESULTS 10-fold CV

Feature set	Max	Average	Min
SMOTE CLEAN	0.921	0.867	0.814
SMOTE ALL	0.927	0.861	0.799
SMOTE REDUCED	0.914	0.827	0.76
CLEAN	0.851	0.861	0.833
ALL	0.85	0.85	0.833
REDUCED	0.846	0.83	0.82

RANDOM FOREST CLASSIFIER MODEL RESULTS ON VALIDATION SET (5-foldCV)

Feature set	ACC	PRECISION	RECALL	F1-SCORE	ROC_AUC	LATENCY
SMOTE CLEAN	0.857	0.609	0.298	0.4	0.631	49.3ms
SMOTE ALL	0.854	0.6	0.255	0.358	0.611	43.2ms
SMOTE REDUCED	0.837	0.471	0.17	0.25	0.567	26.5ms
CLEAN	0.847	0.667	0.085	0.151	0.539	62.5ms
ALL	0.857	0.778	0.149	0.25	0.57	66.5ms
REDUCED	0.837	0.444	0.085	0.143	0.532	175.7ms

RANDOM FOREST CLASSIFIER MODEL RESULTS ON VALIDATION SET (10-foldCV)

Feature set	ACC	PRECISION	RECALL	F1-SCORE	ROC_AUC	LATENCY
SMOTE CLEAN	0.867	0.7	0.298	0.418	0.637	22.8ms
SMOTE ALL	0.861	0.6	0.383	0.468	0.667	19.8ms
SMOTE REDUCED	0.827	0.40	0.17	0.239	0.561	46ms
CLEAN	0.861	0.875	0.149	0.255	0.572	44.2ms
ALL	0.85	0.714	0.106	0.185	0.549	42.5ms
REDUCED	0.83	0.286	0.043	0.074	0.511	7.7ms

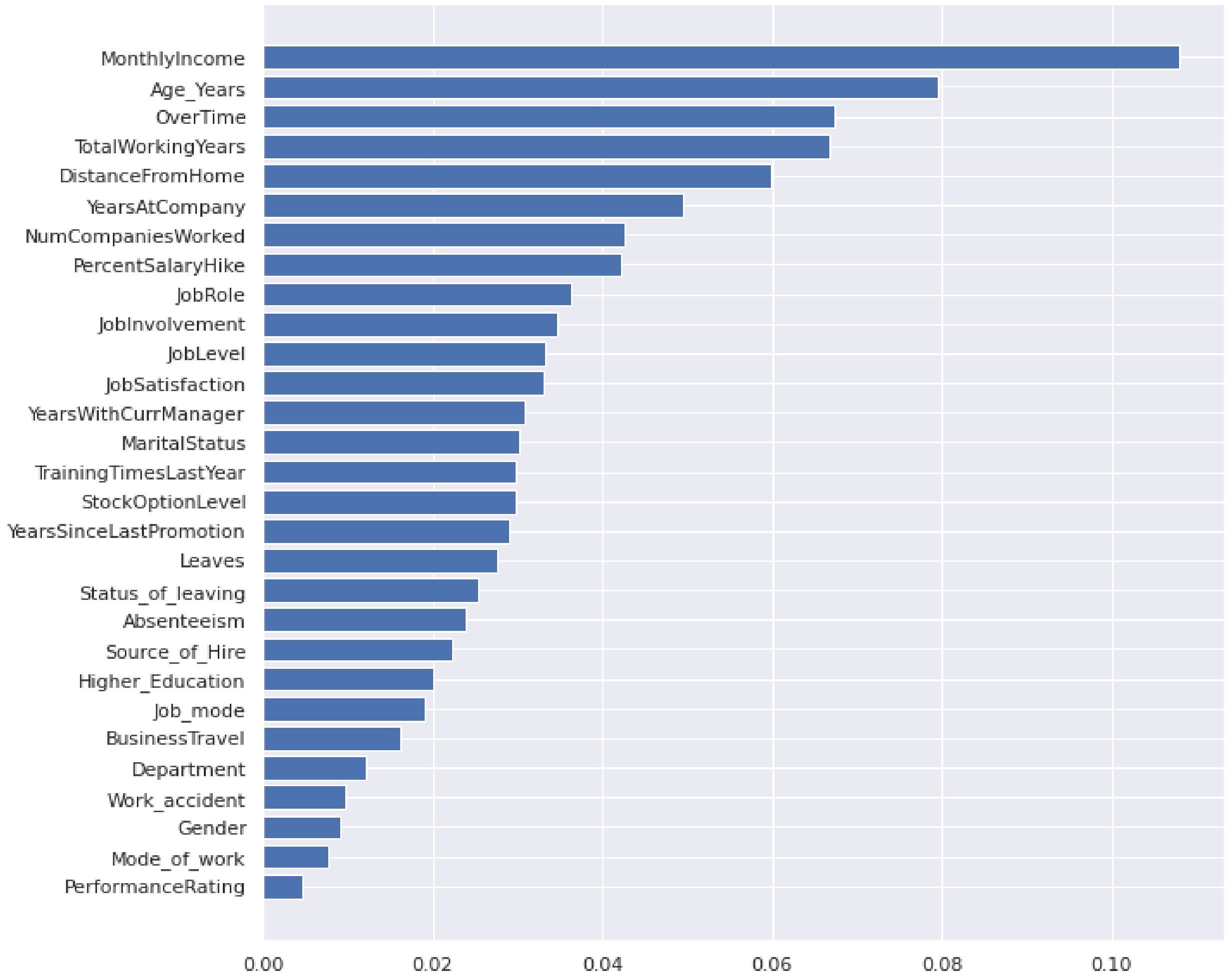
RANDOM FOREST CLASSIFIER MODEL RESULTS ON TEST SET (5-foldCV)

Feature set	ACC	PRECISION	RECALL	F1-SCORE	ROC_AUC	LATENCY
SMOTE CLEAN	0.878	0.652	0.349	0.455	0.658	49.5ms
SMOTE ALL	0.871	0.6	0.349	0.441	0.654	40.6ms

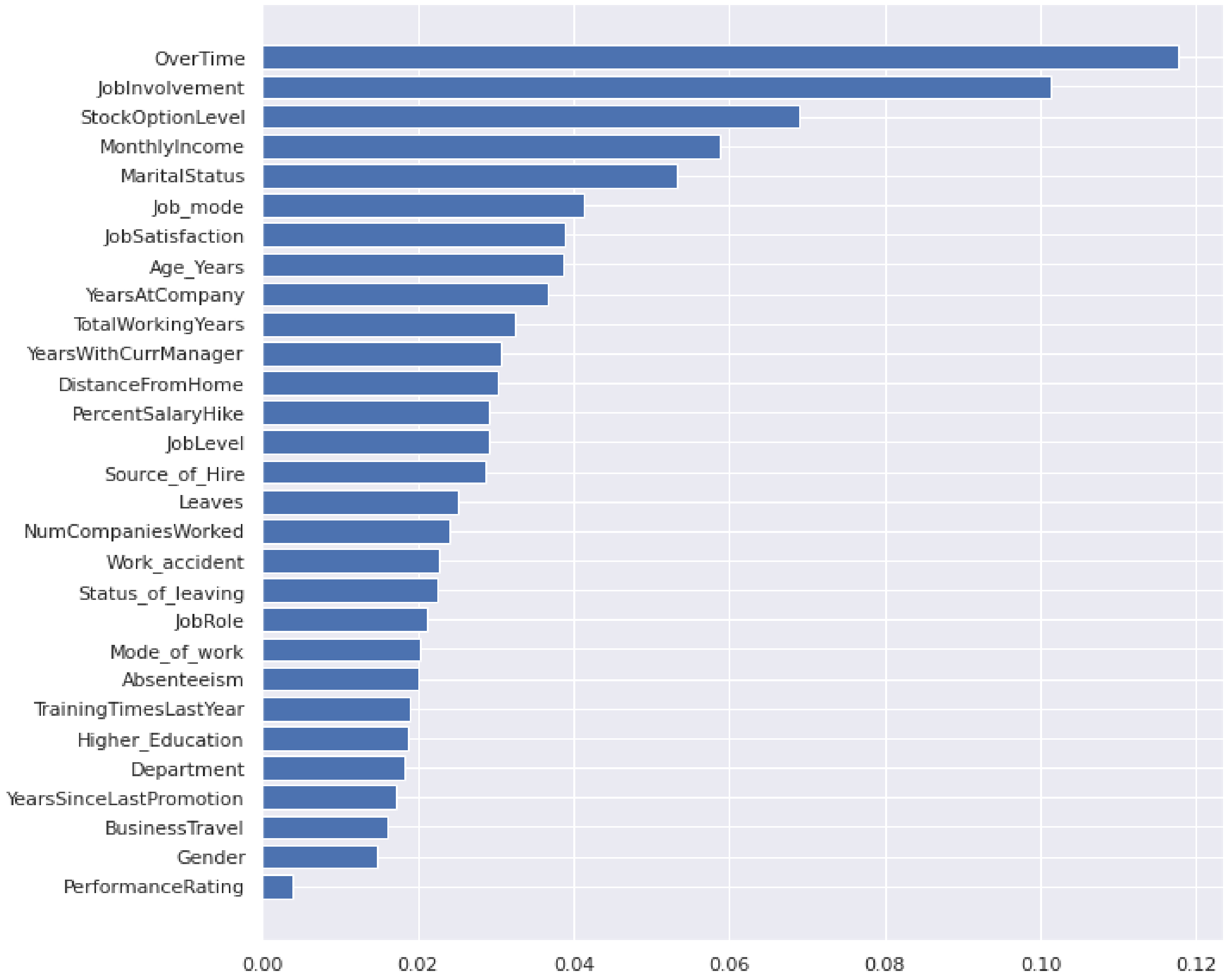
RANDOM FOREST CLASSIFIER MODEL RESULTS ON TEST SET (10-foldCV)

Feature set	ACC	PRECISION	RECALL	F1-SCORE	ROC_AUC	LATENCY
SMOTE CLEAN	0.864	0.565	0.302	0.394	0.631	36.7ms
SMOTE ALL	0.867	0.577	0.349	0.435	0.653	16.3ms

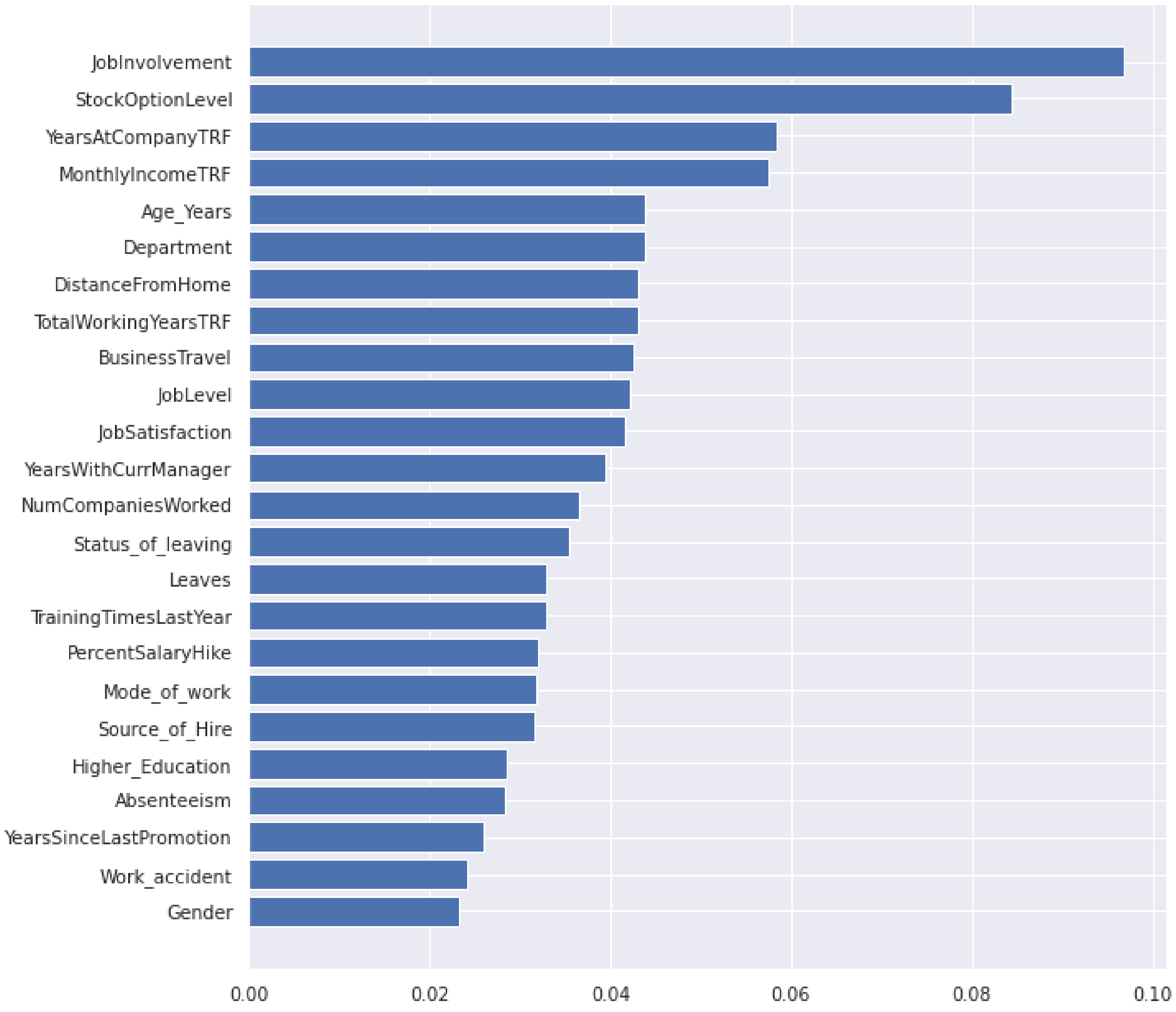
CLEANED FEATURE IMPORTANCE PLOT



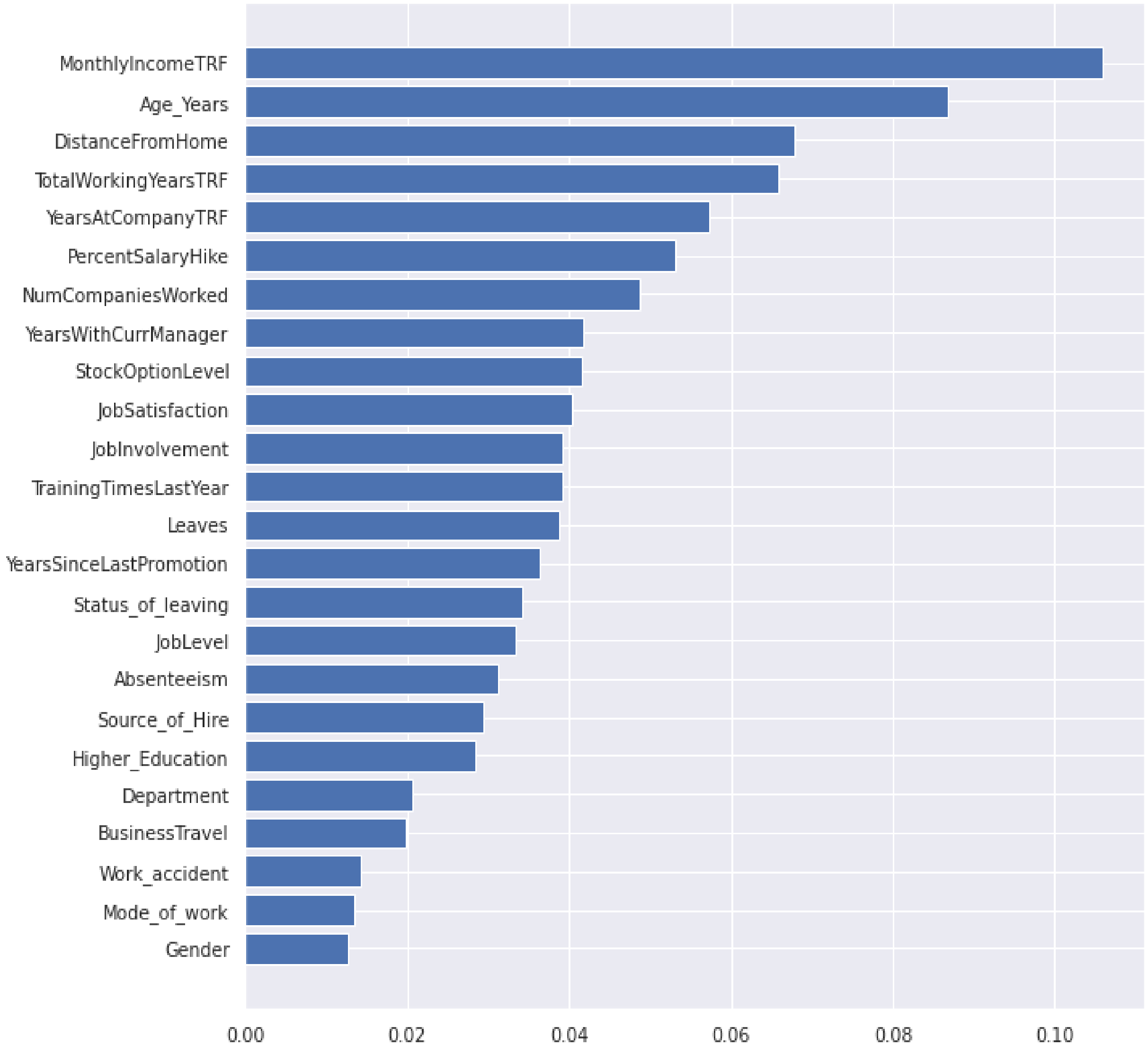
SMOTE CLEANED FEATURE PLOT



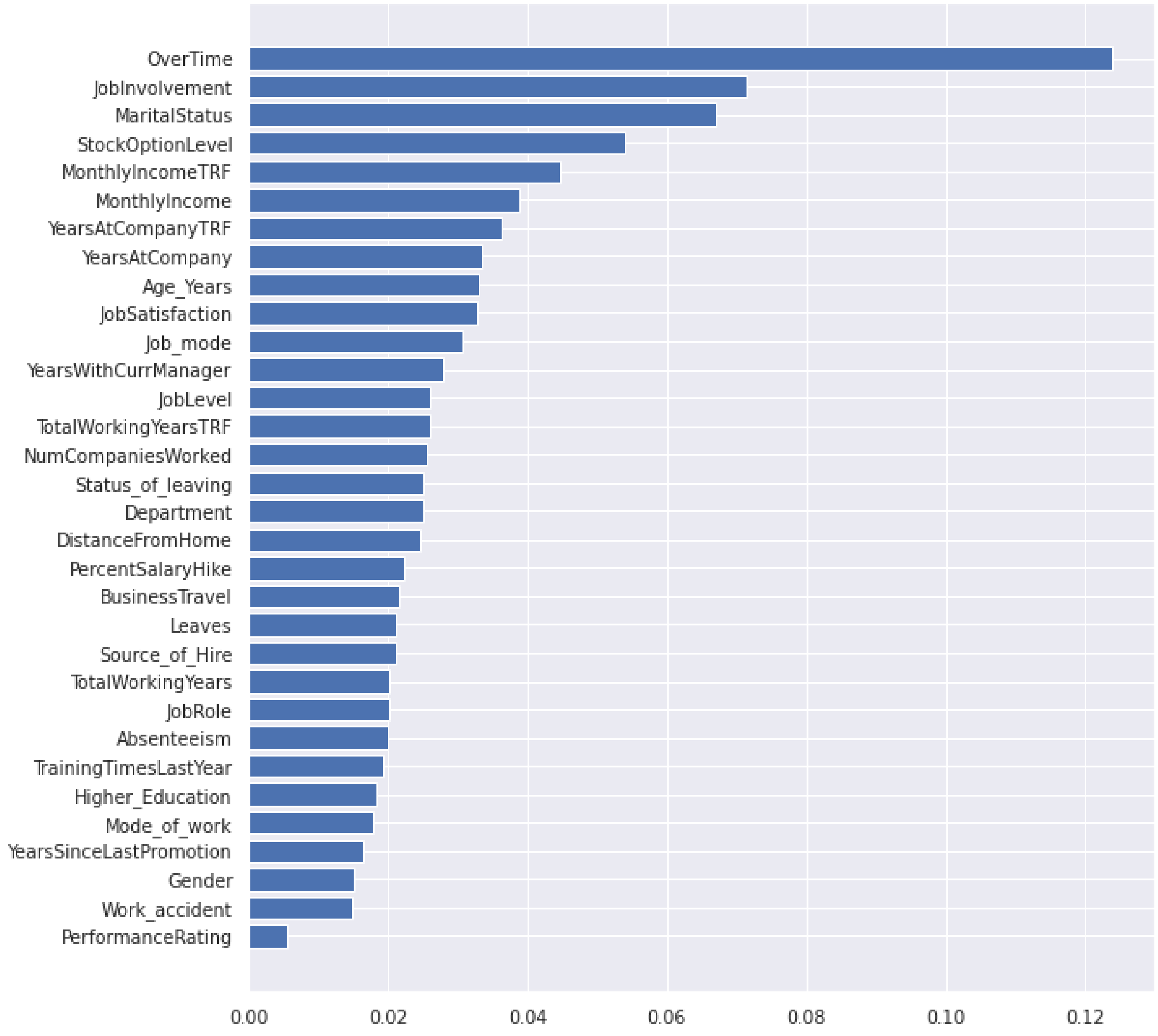
SMOTE REDUCED FEATURE



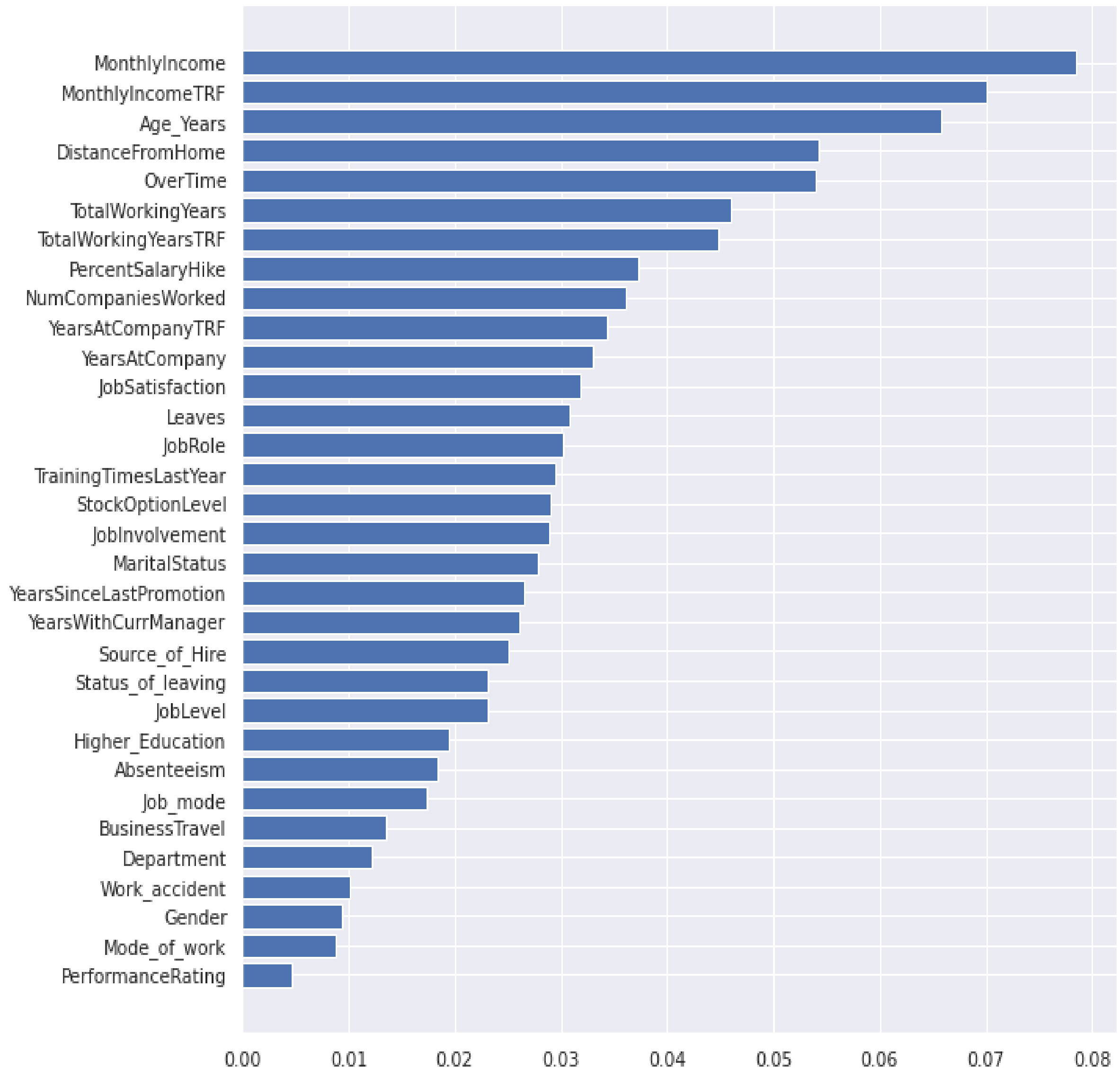
REDUCED FEATURE



SMOTE ALL FEATURE



ALL FEATURE



CONCLUSION

The feature set with the best results in accuracy, precision, recall, f1 score, roc-auc is the set with the SMOTE cleaned features (5-fold cv). However, the SMOTE with ALL features combined did very well with its result very similar to the previous feature set but its computational duration was the smallest.

CHALLENGES, IN PROGRESS & FUTURE WORK

- ❖ I had to create as many checkpoints as possible using pickle so as not to lose unique feature sets
- ❖ The Entire Code's runtime takes quite a while
- ❖ Data imbalance

- ✓ Final Report compilation [one week]
- ✓ Try other classification algorithm to backup effectiveness of combined machine learning technique proposed in this project.

- ❑ Propose ways to further improve the predictive performance of this classification algorithm considering lesser but very influential factors.
- ❑ Propose ways to achieve more accurate prediction with shorter computational runtime
- ❑ Propose other ways the model's hyper-parameter can be tuned to optimize the predictive result of the machine learning algorithm.

REFERENCES

1) NESRINE BEN YAHIA , JIHEN HLEL , AND RICARDO COLOMO-PALACIOS., (Senior Member, IEEE). (2021). "From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction", 1RIADI Laboratory, National School of Computer Sciences, University of Manouba, Manouba , Tunisia Computer Science Department, Østfold University College, 1783 Halden, Norway. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9409047
2) Saeed Najafi-Zangeneh, Naser Shams-Gharneh, Ali Arjomandi-Nezhad, Sarfaraz Hashemkhani Zolfani. (2021). "An improved machine learning-based employee attrition prediction framework with emphasis on feature selection". Mathematics , 9,1226. https://doi.org/10.3390/math9111226
3) Data set title and link: Human resources data based on international business machine corporations (IBM) attrition. https://www.kaggle.com/datasets/singhnproud77/hr-attrition-dataset
4) Ted Hessing. "Box Cox transformation". Six Sigma study guide article. https://sixsigmastudyguide.com/box-cox-transformation/
5) Tanasescu, LG., Bologa, AR. (2022). Machine Learning and Data Mining Techniques for Human Resource Optimization Process—Employee Attrition. In: Ciurea, C., Boja, C., Pocatilu, P., Doinea, M. (eds) Education, Research and Business Technologies. Smart Innovation, Systems and Technologies, vol 276. Springer, Singapore. https://doi.org/10.1007/978-981-16-8866-9_22
6) Shawni Dutta, Samir Kumar Bandyopadhyay. (2020). "Employee Attrition Prediction Using Neural Network Cross Validation Method", International Journal of Commmerce and Management Research. ISSN: 2455-1627; Impact Factor: RJIF 5.22. https://www.researchgate.net/profile/Shawni-Dutta/publication/341878934_Employee_attrition_prediction_using_neural_network_cross_validation_method/links/5ed7becf299bf1c67d352327/Employee-attrition-prediction-using-neural-network-cross-validation-method.pdf
7) Fallucchi, Francesca, Marco Coladangelo, Romeo Giuliano, and Ernesto William De Luca. 2020. "Predicting Employee Attrition Using Machine Learning Techniques" *Computers* 9, no. 4: 86. https://doi.org/10.3390/computers9040086

THANK YOU