C O V E N T R Y
U N I V E R S I T Y

Faculty of Engineering, Environment and Computing

School of Computing, Electronics and Mathematics

MSc. Data Science and Computational Intelligence
EECT044
7151CEM
Computing Individual Project

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

Author: Uchechukwu Fredrick Okonkwo

SID: 11738697

Supervisor: Vernon Sassman

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in Data Science and Computational Intelligence

Academic Year: 2021/22

Declaration of Originality

I declare that this project is all my own work and has not been copied in part or in whole from any other source except where duly acknowledged. As such, all use of previously published work (from books, journals, magazines, internet etc.) has been acknowledged by citation within the main report to an item in the References or Bibliography lists. I also agree that an electronic copy of this project may be stored and used for the purposes of plagiarism prevention and detection.

Statement of copyright

I acknowledge that the copyright of this project report, and any product developed as part of the project, belong to Coventry University. Support, including funding, is available to commercialise products and services developed by staff and students. Any revenue that is generated is split with the inventor/s of the product or service. For further information please see www.coventry.ac.uk/ipr or contact ipr@coventry.ac.uk.

Statement of ethical engagement

I declare that a proposal for this project has been submitted to the Coventry University ethics monitoring website (https://ethics.coventry.ac.uk/) and that the application number is listed below (Note: Projects without an ethical application number will be rejected for marking)

Signed: UCHECHUKWU FREDRICK OKONKWO Date: 18/08/2022

Please complete all fields.

First Name:	Uchechukwu
Last Name:	Fredrick Okonkwo
Student ID number	11738697
Ethics Application Number	P137917
1st Supervisor Name	Vernon Sassman
2nd Supervisor Name	Zina Jerjees

This form must be completed, scanned, and included with your project submission to Turnitin. Failure to append these declarations may result in your project being rejected for marking.

Abstract

Employee attrition is the voluntary or involuntary termination of an employee in an organization, and it is a significant concern for most organization with a functioning HR department as it affects not just their efficiency or work sustainability but also their long-term development tactics. This research work focuses on discussing the combination of various systematic machine learning techniques and statistical data analysis to enhance the predictive accuracy of employee attrition.

The steps taken to achieve this are as follows: data acquisition, exploratory data analysis, data cleaning, feature engineering which includes skewed feature transformation using the box-cox power transformation, determining predictive power of categorical variables using chi-square method for feature selection before encoding all categorical features, applied synthetic minority over-sampling technique (SMOTE) to tackle data imbalance before applying standard scaling and creating a dataset with cleaned features, dataset with reduced feature and dataset with all features. The data split was train-validation-test at 60%-20%-20% respectively. Then I ran 5-fold cross validation on each of the feature sets and selected the best models using the Random Forest classifier, afterwards I evaluated the models on the validation set and picked the best one. Finally, I evaluated that best model on the test set to gauge the model's ability to generalize to unseen data and to confirm its consistency. The evaluation was based on the model's accuracy, precision, recall, f-score, roc-auc and latency or computational speed.

The dataset with clean features and SMOTE applied on the train set gave a maximum accuracy of 91% with max depth of 16 and 32 estimator while the minimum accuracy was 79% with max depth of 2 and 16 estimators this happens to be the best results compared to the other 5 various feature sets. At the validation stage, the accuracy was 85.7%, precision was 60.9%, recall was 30%, f1-score was 0.40, roc-auc was 0.631 and latency was 49.3ms. Finally, at the test stage we had an accuracy of 87.8%, precision of 65.2%, recall of 35%, f1-score of 0.455, roc-auc score became 0.658 and latency of 49.5ms. These results show the influence of the combined techniques in enhancing the accuracy of the attrition class prediction model. Additional analysis was carried out to help the HR department identify the key factors that influence attrition which could help them in minimizing it and enhance retention on the other hand.

Table of Contents

Abstract 2
_____ 2
Table of Contents 3
Acknowledgements 5
1 Introduction 6
1.1 Problem Statement 6
1.2 Project Aim and Objectives 6
2 Literature Review 8
2.1 Machine learning binary classification 8
2.2 FEATURE ENGINEERING 10
2.3 SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE) 12
3 Methodology 14
3.1 DATASET 14
3.2 EXPLORATORY DATA ANALYSIS(EDA) AND TABLEAU VISUALS 16
3.3 Data Pre-processing 21
3.4 Modelling 24
4 Techniques 28
4.1 Feature Selection 28
4.2 Synthetic minority over-sampling technique (SMOTE) 29
4.3 Feature Engineering 30
5 Result and Discussion 32
6 Project Management 36
6.1 PROJECT SCHEDULE 36
6.2 RISK MANAGEMENT 36
6.3 QUALITY MANAGEMENT 36
6.4 Social, Legal, Ethical and professional considerations 37
7 Conclusion 38
7.1 Achievements 38
7.2 Future work 38
8 Student reflections 38
8.1 Achievements ..... Error! Bookmark not defined.
8.2 Future Work ..... Error! Bookmark not defined.
Bibliography and References 39
Appendix A – Project Proposal 1

Appendix B – Project Presentation 1

Appendix C – Meeting records 1

Appendix X – Visualizations 1

Figure 1 Performance evaluation of models on the real dataset. N.B Yahia et al ----------9
Figure 2 performance metrics results. Francesca et all.2020 ---------- 10
Figure 3 feature selection steps as proposed by N.B Yahia et al ---------- 11
Figure 4 classification accuracy obtained by the various feature selection methods. M. Subhashini et al. 2020---------- 11
Figure 5 Precision score obtained by various feature selection methods using different classifiers. M. Subhashini et all.2020 ---------- 12
Figure 6 illustrates the different over-sampling approaches. (Cheng Zhang et al. 2020) - 13
Figure 7 Result of the analysis (Cheng Zhang et all. 2020) ---------- 13
Figure 8 Correlation heatmap showing relationship between continuous features -------- 16
Figure 9 Distribution of the continuous variables ---------- 17
Figure 10 Attrition vs number of companies worked and distance from home. (Tableau dashboard 1) ---------- 18
Figure 11 Attrition vs Date hire, total working years and years at work. (Tableau Dashboard 2)---------- 18
Figure 12 Attrition vs years since last promotion, salary hike, performance rating, and job satisfaction. (Tableau dashboard 3) ---------- 19
Figure 13 Attrition count vs gender, department, and age group. (Tableau dashboard 4) 19
Figure 14 Attrition vs Marital status, mode of work, job mode, and jobe role. (Tableau dashboard 5) ---------- 20
Figure 15 Attrition vs education level, status of leaving, work accidents, overtime, and source of hire. (Tableau dashboard 6)---------- 20
Figure 16 Box cox transformation of monthly income ---------- 22
Figure 17 Box cox transformation of total working years ---------- 23
Figure 18 Box cox transformation of total years at company ---------- 23
Figure 19 Binary classification---------- 24
Figure 20 A typical ensemble learning based technique ---------- 25
Figure 21 shows the kth-fold cross validation scheme ---------- 26
Figure 22 logistic regression ---------- 26
Figure 23 SMOTE operation---------- 29
Figure 24 After smote class tends to be balanced---------- 29
Figure 25 feature engineering work flow---------- 30
Figure 26 shows the cleaned feature importance plot with and without SMOTE ---------- 34
Figure 27 shows the reduced feature importance plot with and without SMOTE---------- 34
Figure 28 shows the reduced feature importance plot with and without SMOTE---------- 35
Figure 29 Gantt chart show project plan and schedule ---------- 36

Acknowledgements

I want to acknowledge Almighty God for his loving kindness and infinite mercy in my live. I also want to appreciate my family members who supported me through this academic journey. My sincere gratitude is extended to my lecturers, as well as my ever-guiding supervisors.

1 Introduction

HR analytics is a multi - disciplinary approach that incorporates methodology to improve the quality of people-related decisions to improve organisational performance. Talent analytics, people analytics, and workforce analytics are all interchangeable terms for HR analytics. HR analytics is used in all components of the human resources function, including recruiting, training and development, succession planning, retention/attrition, engagement, compensation, and benefits (Sujeet et al, 2016). Human resource analysts are those that employ the use of "high-end" predictive modelling to forecast the effects of changing policies or conditions.

Traditional HR analytics is preoccupied with what is going on in the present, for instance, issues related to turnover and cost per hire. However, most institutions lacked a consistent and comprehensive view of their workforce, hence they needed to necessitate the use of HR analytics to optimize work force performance. At some point, it became essential for HR to develop IT and finance analytical skills and capabilities to achieve a better return on investment.

Evolution of technological procedures and techniques when amalgamated with predictive analytics gave way to an enhancement of HR purposes in the last few decades. This combination of effective techniques generates valuable insights that could never have been achieved through the traditional methods, which we must agree is a reactive and evidence-based system whereas the HR predictive analytics is proactive and has a broader fact-based decision system.

1.1 Problem Statement

Attrition prediction is essentially predicting the employees that are most likely to quit their job based on some factors like Age, marital status, general performance, job satisfaction, absenteeism, years spent with current company, monthly income, distance of residence to work, mode of work and so on. From the company's point of view, employee attrition is a significant metric in HR analytics. It is essential to gain this information because recruiting employees is often arduous and costlier than retaining old ones. It also necessary because when trained, technical and resourceful employees leave a place of employment for another job which offers better opportunities, this often results in financial loss and low morale of remaining employees. 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. Therefore, companies make use of this information to see how they can prevent the employee from leaving by improving the employee retention.

Those who stand to benefit maximally from this project are the human resources managers as well as the project managers. Applying this comparative machine learning predictive analysis techniques on the employee's data gives the HR managers control over the attrition rate saving the company a lot of cost and resources. The general idea is to improve retention rate. However, to achieve this, careful decision must be made on which employees are valuable and profitable to the organization so they can be retained.

1.2 Project Aim and Objectives

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:

- Model Accuracy: percentage of the total correct prediction over total number of examples.
- Model Precision: percentage of the total number employees the model predicted to attrite that churned over the total number of employees that the model predicted to attrite, that is, when the model predict, what percentage of the time was the model correct? How many selected items are relevant?
- Model Recall: this tells the percentage of how many relevant items are selected by the model.
- F1 score: this is a statistical measure that calculates the harmonic mean between precision and recall respectively
- ROC_AUC
- Latency (computational time)

We will be addressing questions like; will the application of this combined techniques which include advance feature engineering, hyperparameter tuning, multiple feature sets, cross validation, and synthetic minority oversampling technique (SMOTE) to improve the accuracy, precision and recall of the model's class prediction or not?

Additional aim is to identify the most influential factors that affect the attrition of employees.

2 Literature Review

Employee attrition or voluntary turnover is a significant concern for most organization with a functioning HR department as it affects not just their efficiency or work sustainability but also their long-term development tactics. For this reason, many researchers have employed various techniques into people analytics to comprehend and explain employee attrition so they could recommend solutions to help human resource managers to reduce attrition by improving the way they recruit and retain resourceful employees. The task is predictive analysis which is a branch of advanced analytics that makes predictions about future outcomes based on historical data combined with statistical modelling, feature engineering techniques and machine learning as a synergy. Organizations employ predictive analytics to find patterns in the data to identify problems, opportunities, and underlying circumstances, hence, proffer specific solutions centred on the analysis outcome.

2.1 Machine learning binary classification

Research reports show several employee attrition or voluntary turnover classification predictive models as well as various approach to techniques with the major aim to enhance the model performance metrics. For this binary class prediction task, researchers have used logistic regression, decision tree, support vector machines, K-nearest neighbour, naïve bayes, XGboost and random forest classifier with a varying combination of dimension reduction (principal component analysis), feature selection, model optimization, ensembling methods and SMOTE for data imbalance problems.

Nesrine, Jihen and Ricardo in 2021 used the following classifiers: Decision Tree, Logistic Regression and Support Vector Machine (as machine learning models), Random Forest, XGBoost and Vote Classifier (as ensemble learning models) and three deep learning models (DNN, LSTM and CNN). A grid search algorithm was performed for each classifier over tuning hyperparameters. Then, the different models were trained using their best configuration on the training dataset. They evaluated their results, and it showed the outcome was improved after two feature selection techniques, Recursive Feature Elimination (a wrapper method) and SelectKBest (a filter method), was applied in the analysis. In this method, they began by thinking big by collecting most of the common features from the literature (an exploratory study), and then they attempted to think 'deep' by filtering and selecting the most important features using a survey and feature selection. When compared to the previous solutions, their approach achieves higher accuracy (0.96, 0.98, and 0.99, respectively) for the three datasets.

Finally, while rewards and payments are commonly thought to be the most important factors in employee retention, their findings show that 'business travel,' which is less popular in the literature, is the leading influence for employees and should be included in HR policies to ensure retention.

The machine learning based predictive models used are the Decision tree(DT), Support vector machine (SVM), and logistic regression(LR). The deep learning-based approach was the deep neural networks(DNN), long short-term memory networks (LSTM), convolutional neural networks(CNN). While the in the ensemble learning based predictive models, they used the

random forest, XGBoost, Voting classifier (VC)that trains the classifiers and then predicts the target class on a majority vote based on the hard voting and the soft voting strategy. The stacked ANN-based model is the last model used in the ensemble learning based models where outputs from the above three deep learners are collected to make a new dataset.

Models	IBM dataset		11 features from IBM dataset		Kaggle dataset	
	Acc.	F1	Acc.	F1	Acc.	F1
DT	0.777	0.318	0.952	0.766	0.972	0.945
LR	0.83	0.368	0.849	0.15	0.782	0.37
SVM	0.85	0.5	0.837	0.42	0.782	0.495
DNN	0.80	0.42	0.84	0.4	0.89	0.18
LSTM	0.71	0.487	0.75	0.57	0.79	0.605
CNN	0.84	0.672	0.89	0.649	0.91	0.7
RF	0.858	0.169	0.953	0.828	0.978	0.967
XGB	0.853	0.434	0.956	0.728	0.976	0.946
VC	**0.93**	0.58	**0.96**	0.62	**0.98**	0.88
Stacked	0.88	0.5	0.9	0.631	0.96	0.67

Figure 1 Performance evaluation of models on the real dataset. N.B Yahia et al

The table shows that the ensemble learning based predictive models had the best performance. The voting classifier had the best accuracy score of 0.98.

It is critical to select the algorithms appropriately to correctly predict an employee's behaviour and reach a conclusion. Classifications in machine learning can have two different meanings. On the one hand, we may be given a set of observations with the goal of determining the presence of various clusters and classes in the data. On the other hand, we may be certain that there are a specific number of classes.

Francesca et al.2020 and her team had a goal to analyse how objective factors affect employee attrition to identify the main causes that contribute to a worker's decision to leave a company and to be able to predict whether a specific employee will leave the company. Following training, the obtained model for predicting employee attrition is tested on a real dataset provided by IBM analytics, which includes 35 features and approximately 1500 samples. The results are expressed in terms of classical metrics, and the algorithm that produced the best results for the available dataset is the Gaussian Nave Bayes classifier. It reveals the best recall rate (0.54) since it describes a classifier's ability to find all positive instances and achieves an overall false negative rate equal to 4.5% of the total observations as shown in the figure 2 below. The modelling procedure involved selecting models from various machine learning techniques used in their experiment. The classification algorithms considered as as follows:

- Gaussian Naïve Bayes
- Naïve Bayes classifier for multivariate Bernoulli models
- Logistic regression classifier
- K-nearest neighbours (KNN)
- Decision Tree

- Random forest
- Support vector machine classifier (SVM)
- Linear support vector machine (LSVM) classifier

	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

Figure 2 performance metrics results. Francesca et all.2020

From this research work we realise that for us to properly evaluate an algorithm's performance, we need not focus on just the accuracy score as the overall basis for conclusion. In this case, the recall score is considered since it reveals a model's ability to correctly predict the positive instances as it is in the real data.

2.2 FEATURE ENGINEERING

This is the pre-processing step of which involves the selecting, manipulating, and transformation of raw data into features, using statistical and machine learning approaches, that can be used for creating a predictive model in supervised learning (Hashil Patel. 2021). It is closely related feature selection, dimension reduction, feature extraction, exploratory data analysis and so on. They are all involved in the process of designing artificial feature into an algorithm in order to improve its performance without affecting in any way the information the raw data held before the modifications.

N.B Tahia et al (2021) employed the use of the exploratory study method as a first approach by relying on the secondary research (past related works) which helped identify the features that had most impact on solving the problem related to employee attrition. They went further to make use of two popular feature selection method s which are, Recursive feature elimination (RFE) which is an example of the wrapper method as well as the selectkbest which is a filter method. Their steps are depicted in the figure 3 below

Finally, based on the combination of the two feature selection techniques (RFE and SelectKBest) and the data collected, the 11 main features required for employee attrition prediction are as follows: Age, Marital status, Tenure, Grade, Rewards, Job involvement, Training, Business Travel, Job satisfaction, Job performance, and Environment satisfaction.

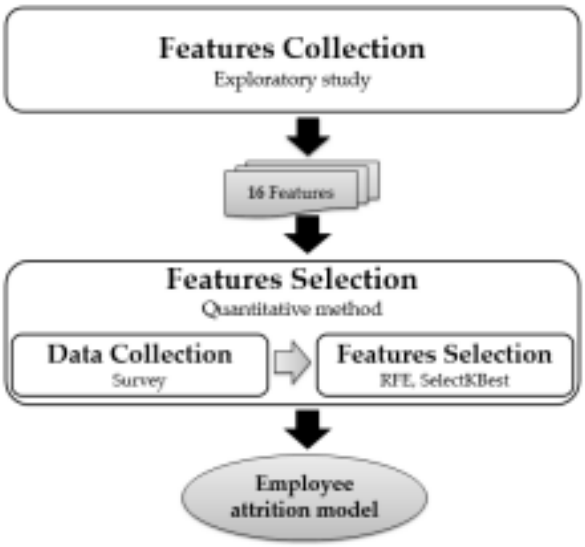


Figure 3 feature selection steps as proposed by N.B Yahia et al

With focus on feature engineering, Dr M. Subhashini et al (2020) and his co-researcher employed the feature selection strategy to improve their classification accuracy, precision, and recall respectively. They used various effective feature selection techniques, such as information gain, which considers the most significant function with a high information gain as the most relevant function, but the fundamental disadvantage of this techniques is that it chooses a high-data-gain feature that may be more informative or otherwise, and this knowledge acquired is limited in dealing with redundant features. The gain ratio feature selection is a revised information gain version which takes in consideration daughter nodes which are nodes where a feature separates the class distinction data. The chi-square feature selection method majorly addresses the categorical columns by checking the predictive power of this attributes based on the best chi-square score or lowest p- values. The correlation-based feature selection is used to address the numerical feature which could be the spearman rank correlation plot or the persons correlation plot which show the relationship between each of the numerical variables in the feature set. The correlation coefficient equal to or close to 1 indicates a strong positive relationship while a correlation coefficient equal to or close to 0 indicates a strong negative relationship between the two features under consideration. While the fisher exact test uses the famous Fisher score to select the appropriate attributes using the Fisher ratio concept and a pattern recognition approach to establishing a value for qualities. The effect of this feature selection methods on the model's performance are shown in the figure 4 and 5 below.

Feature Selection Methods	Classification Accuracy (in %) by Classification Techniques					
	ANN	SVM	GBT	Bagging	RF	DT
Original Dataset	44.27	45.31	49.21	43.98	41.65	42.86
Correlation based FS	68.51	68.76	69.73	67.31	66.82	65.41
Information Gain	64.33	65.74	69.81	62.23	63.71	65.43
Gain Ratio	64.43	64.74	66.64	64.34	61.82	62.77
Chi-Square	89.44	85.75	89.15	79.81	80.53	81.22
Fisher's Exact	72.65	73.21	73.76	68.92	69.16	70.88

Figure 4 classification accuracy obtained by the various feature selection methods. M. Subhashini et al. 2020

It is clear from the results that the chi-square feature selection method had the greatest impact on the model's performance in predicting the employee's attrition.

Feature Selection Methods	Precision (in %) by Classification Techniques					
	ANN	SVM	GBT	Bagging	RF	DT
Original Dataset	44.72	48.16	50.61	43.43	44.31	45.76
Correlation based FS	67.67	67.52	72.51	66.97	65.43	66.81
Information Gain	58.39	57.61	60.43	55.64	54.32	59.45
Gain Ratio	57.68	57.52	59.52	54.43	52.16	58.65
Chi-Square	83.18	85.24	85.42	81.53	79.72	78.18
Fisher's Exact	55.85	56.21	61.28	65.89	64.25	55.32

Figure 5 Precision score obtained by various feature selection methods using different classifiers. M. Subhashini et all.2020

2.3 SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE (SMOTE)

In past years, classification of imbalanced data has affected most machine learning class prediction performances in a negative way. Due to this imbalance nature of datasets, the algorithm tends to be biased, that is, its predictions always favour the majority target class of which in most cases is not the class under consideration or of interest. In other to address this issue and hereby focus on the target of interest, which in most case is the minority class, various methods have been employed depending on the aim of the analysis (Cheng et al.2020).An over-sampling method was proposed in Cheng Zhang and his research mates for imbalanced data classification called abstention-SMOTE. Firstly, they created abstaining classifiers using ROC analysis, then used the abstaining classifiers to generate the abstention positive samplings, which only include positive samples that are easily misclassified. Finally, we use these positive abstention samples to create new positive samples to balance the data distribution. Unlike existing over-sampling methods, where the over-sampling of all minority examples or a random subset of the minority class is done, they only over-sample the abstention minority class. Different shapes of points in two-dimensional space represent different classes of samples, such as negative samples, positive samples. Some of the synthetic data generated by RO-sampling and SMOTE are far from the minority class, as we can see. This is due to the fact that the distribution of data sets is not taken into account, and these samples contribute less to expanding the minority class border, even affecting the classifier's performance. The synthetic instances generated by Borderline-SMOTE are close to the minority class border, however, some are far from the minority class, which will change the data distribution and have a detrimental impact on the class prediction model. For this reason, they proposed a technique that generates synthetic instances which are similar to minority class samples that are easily misclassified. This way they can contribute more to the expansion of the minority class border no matter how dispersed they are. The different over-sampling methods are illustrated in figure 6 below.

Input: Testing set and training set, the cost coefficients c_{12}, c_{21}, the fraction k_{max} and the nearest neighbors t.

Output: Classification results of the test set.

Step1. Input the training set and testing set into Naive Bayesian classifier and collect statistics to construct ROC curve.

Step 2. Search the $(fp_\alpha, tp_\alpha), (fp_\beta, tp_\beta)$ to construct the optimal abstaining classifier $A_{\alpha\beta}$ and get the positive abstention samples.

Step 3. For each positive abstention sample, they calculated its nearest neighbors *t* from the positive abstention sample and generate *t –1* new positive instance.

Step 4. Randomly select some new positive instances to balance the training data, then use the new training set to train our classifier and get the classification results of the testing set.

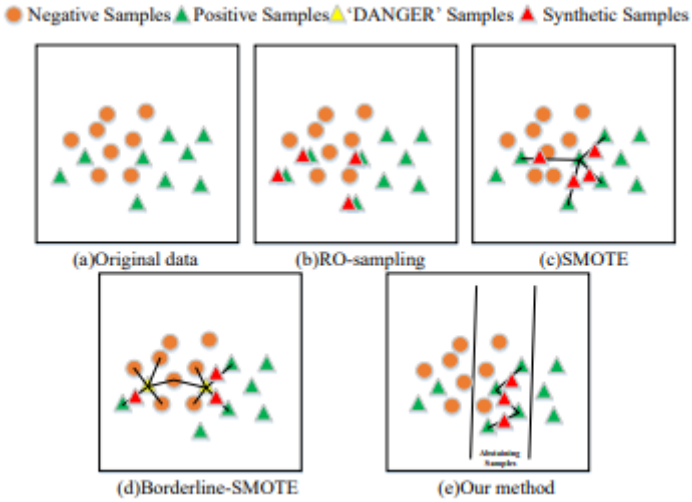


Figure 6 illustrates the different over-sampling approaches. (Cheng Zhang et al. 2020)

The results of the experiment show that their approach outperforms three other over-sampling methods, namely RO-sampling, SMOTE, and Borderline-SMOTE and this is seen in the figure below showing the average results of the various SMOTE methods applied on the four different data sets.

Data set	Algorithm	Recall (%)	F1 (%)	cost
FLL	NBC	63.33	60.80	0.2906
	RO	66.67	60.15	0.2906
	SM	70.00	62.69	0.2687
	BS	70.00	63.16	0.2666
	Ours	81.67	74.24	0.1750
Heart	NBC	67.50	77.78	0.2296
	RO	75.00	80.36	0.1926
	SM	74.17	80.91	0.1926
	BS	75.00	81.08	0.1889
	Ours	78.33	82.10	0.1722
Breast-w	NBC	97.49	94.91	0.0271
	RO	98.74	95.55	0.0205
	SM	99.16	94.80	0.0220
	BS	99.58	95.01	0.0198
	Ours	99.58	95.58	0.0176
Credit	NBC	42.33	47.12	0.3155
	RO	55.33	53.72	0.2770
	SM	57.67	51.49	0.2900
	BS	61.67	55.22	0.2650
	Ours	62.67	56.04	0.2595

Figure 7 Result of the analysis (Cheng Zhang et all. 2020)

3 Methodology

This chapter addresses the methodology applied in my project based on the related and aim of the project. The best methods are described as well as the reason of choice which is expected to have great impact on the performance of the model. A detail view of the dataset, exploratory data analysis and tableau visualizations, the data pre-processing, modelling, evaluation metrics, hardware and implementation, and finally the machine learning algorithms used, are all presented in this chapter.

3.1 DATASET

One of the challenges every upcoming data scientist face is choosing the right data set that best gives you an opportunity to achieve your aim and objectives for the project's experimental setup or analysis. The dataset was gotten Kaggle website which is an online public data repository site. It is a human resources attrition data set based on the IBM attrition. It consists of 32 columns or features with 1470 inputs belonging to 1470 employees and 17 of these features are integers, that is, numerical columns, while 13 other columns are categorical features.

Table 1 showing dataset description and type

Features	Description	Type
1) Age	Age of employees	Integer
2) Attrition	Yes = attrite, No = never attrite	String(categorical)
3) Business travel	if the employee does not go on business trip, goes frequently or goes rarely	String
4) Department	Human resource, research& development or sales	String
5) Distance from home	Distance from home to work in miles	Integer
6) Gender	Male or female	String
7) Job Involvement	Level of Involvement in job (1-5)	Integer
8) Job level	Job level (1-5)	Integer
9) Job role	Employee's Job role	String
10) Job satisfaction	Level of employee's job satisfaction	Integer
11) Marital status	Employee marital status Single, married or divorced	string
12) Monthly income	Employees monthly income in dollars ($)	Integers
13) Number companies worked	Number of companies employee previously worked for (0-9)	Integers
14) Overtime	Yes = worked overtime, No = didn't work overtime	String (categorical)
15) Percent salary hike	Percentage increase in	integers

	employee salary (%)	
16) Performance rating	Performance rating (3 or 4)	Integers
17) Stock option level	Level of employee stock option (0-3)	Integers
18) Total working years	Total years of working including with previous company (0 – 40)	Integers
19) Training times last year	How many times employee was trained in previous years in company. (0 – 6)	Integers
20) Years at company	Total years spent with present company. (0 – 40)	Integers
21) Years since last promotion	Years since employee was last promoted. (0 - 15)	Integers
22) Years with current manager	Total years employee spent with current manager. (0 - 17)	Integers
23) Higher education	Educational level of employee. 12th grade, BSc, post graduate or PHD	String
24) Status of leaving	Reason for leaving company. Better opportunity, department head, salary, work accident, work environment.	String
25) Mode of work	Employees mode of work. Work from home or work in office.	String
26) Leaves	Number of leaves taken so far. (0 - 5)	Integers
27) Absenteeism	Number of times employee was absent at work. (0 - 3)	Integers
28) Work accidents	Yes = employee had accident at work, No = no accident at work	String (categorical)
29) Source of hire	Through what source did employee get hired. Job event, job portal, recruiter, or walked in	Strings
30) Job mode	Employee's job mode. Contract, full time, or part time	String
31) Date of hire	Date employee was hired	Date
32) Date of termination	Date employee's employment was terminated. (empty)	Date

3.2 EXPLORATORY DATA ANALYSIS(EDA) AND TABLEAU VISUALS

In statistics, exploratory data analysis is a method of analysing data sets to summarise their main characteristics, which is frequently done using graphs and charts and other data visualisation methods. A statistical model may or may not be used, but the primary goal of EDA is to investigate what the data tells us prior to formal modelling or advanced analysis. This is the first leg to finding answers to the problem related questions even though its just a scratch on the surface of analysis. However, this first task could give an overview of what to do or not and how best we can approach the task with the aims and objective in mind.

PEARSON CORRELATION COEFFICIENT

It is principally a standardised measurement of covariance, with the result always falling between -1 and 1. As with covariance, the measure can only reflect a linear correlation of numerical variables and chooses to ignore many other types of associations or correlations. As seen in figure 8 below, the correlation coefficient close to 1 signifies a strong relationship between both features while correlation coefficients close to -1 indicates a very low relationship exist between the two features.

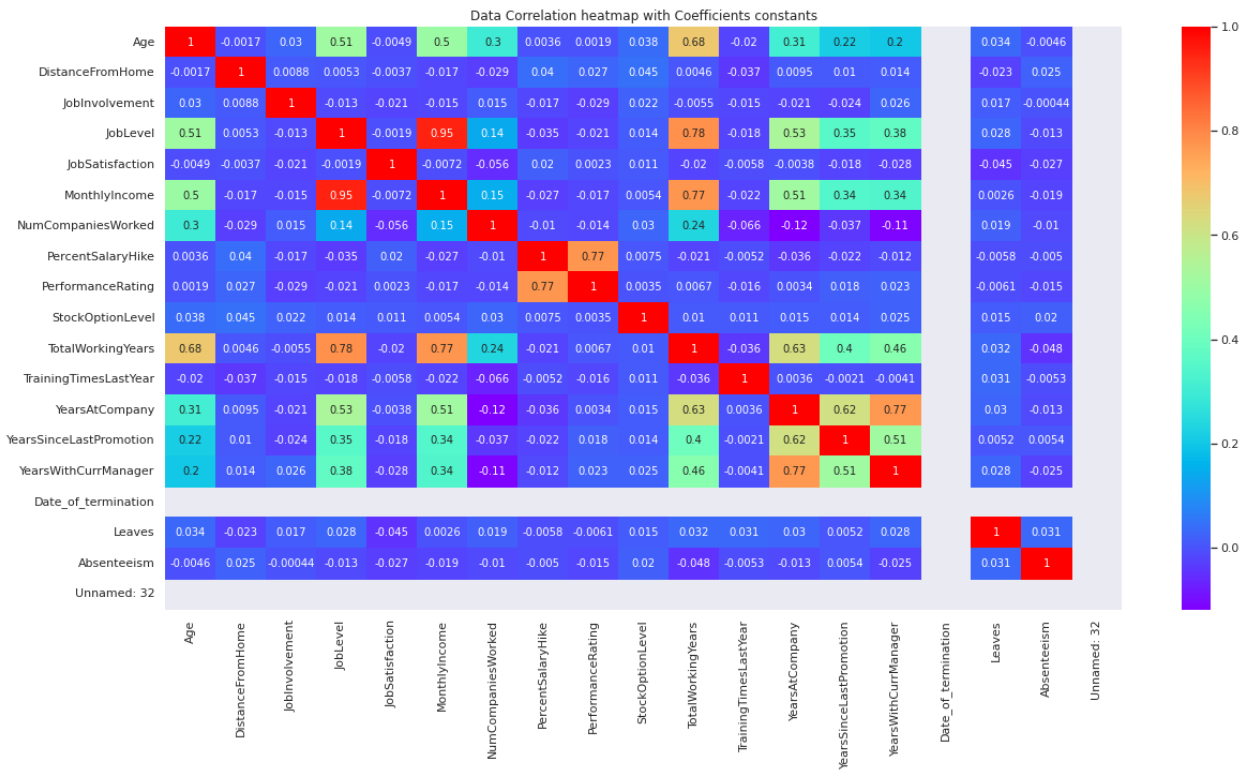


Figure 8 Correlation heatmap showing relationship between continuous features

We notice that the highest correlation coefficient is 0.78 which is the positively strong correlation existing between total working years and job level which is reasonable because an employee who has worked for long would obviously have his/her job level increased over that period. We also notice a strong correlation coefficient of 0.77 that exist between total working years and monthly income, years spent with current manager and years spent in company. This

result validates the credibility of this data and this confirms that it holds useful information that could be used to identify some trend patterns in the data set.

Next is plotting the distribution of continuous variables. These histogram helps us see how distributed these continuous features are. Do they take a normal distribution? Are they right skewed or left skewed? Answers to these questions is very important especially when considering transforming our data in the pre-processing stage.

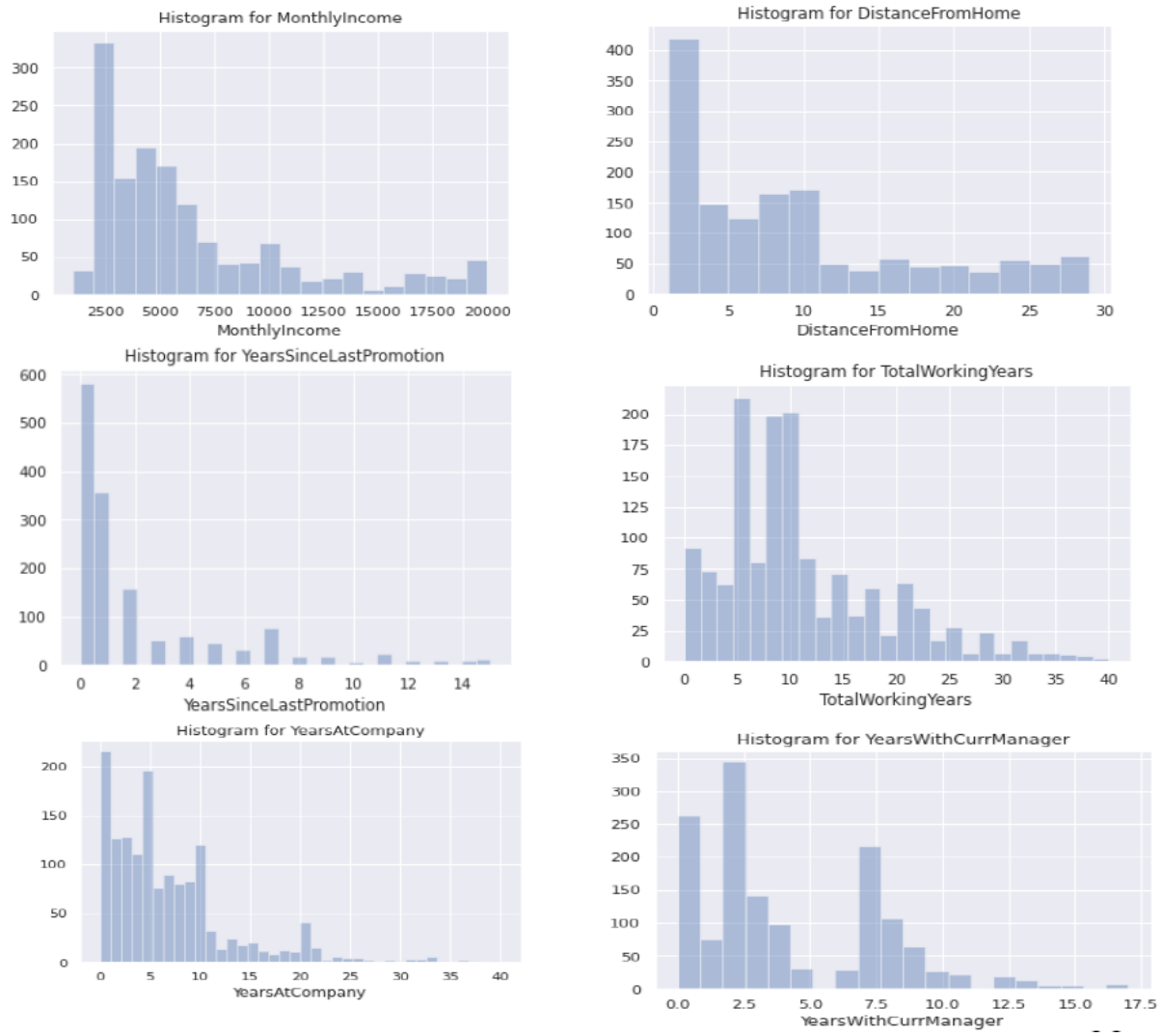


Figure 9 Distribution of the continuous variables

From the figure 9 above which shows the distribution of six continuous feature, monthly income, distance from home, years since last promotion, total working years, years at company, and finally years with current manager, we notice that most of them are right skewed which indicates that the mean of this continuous features is greater than the mode, that is , most of the values are clustered around the left tail of the distribution while the right tail of the distribution is longer.

VISUALIZATIONS USING TABLEAU

I created about 24 interesting visuals, which includes the pie chart, bar chart, line graph for time plot, showing the relation between the count of attrition vs other features both numerical and categorical to see which feature truly influences the attrition. The data was not altered in anyway. I created a new age column that shows the age groups of 18 -25, 26 – 30, 31 – 40, 41 -

50, and 51 – 60. This way I could tell what age group attrite more and further carefully examine other plots to determine what reasons are responsible for the employee's attrition. This sets the motion to discovering answers to the research questions raised at the onset of this project.

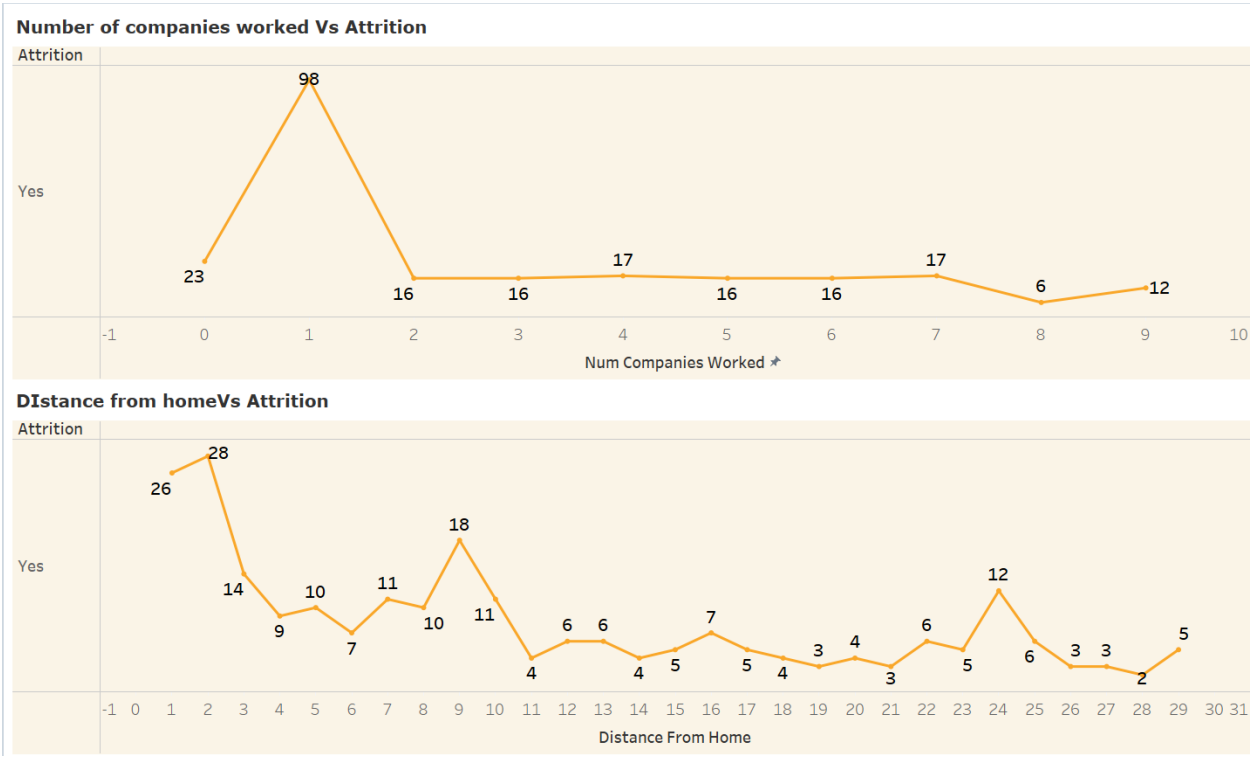


Figure 10 Attrition vs number of companies worked and distance from home. (Tableau dashboard 1)

In figure 10 above we see that employees that has worked for less than 2 companies tend to attrite more as well as those who don't leave far from work. This trend raises a lot of questions about why this happens and the human resource manager could make use of this information to dig deeper into improving employee retention.

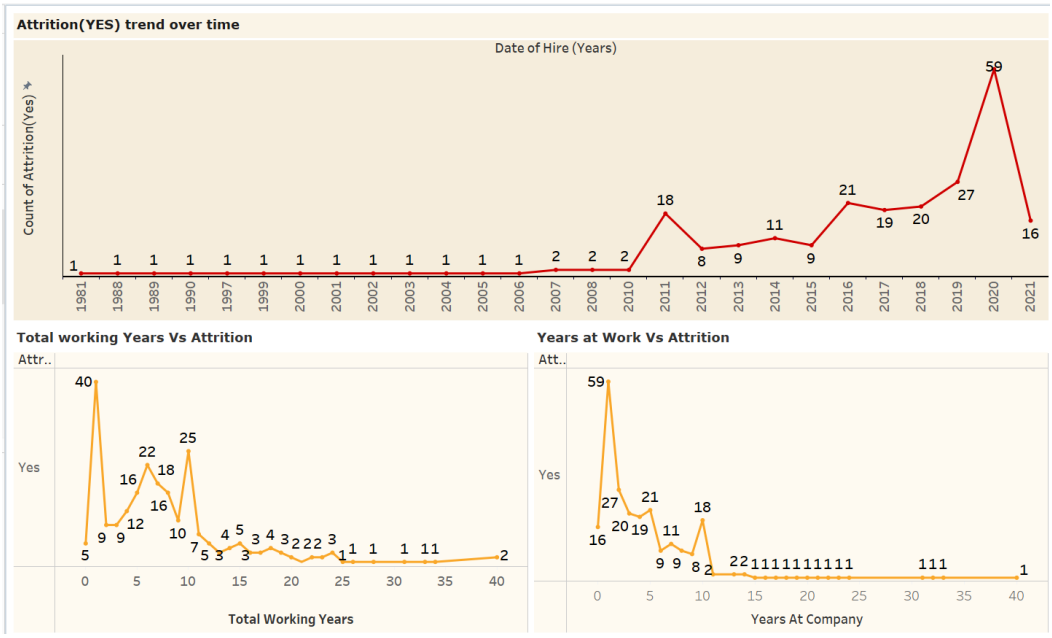


Figure 11 Attrition vs Date hire, total working years and years at work. (Tableau Dashboard 2)

In figure 11 above we see the number of those that left their job at a particular year. It shows that more employees quitted more in 2020 with an all-time number of 59employees followed by 21 in 2016. This prompts the HR manager to trace back what possibly might have caused this high rate of attrition in these respective years.

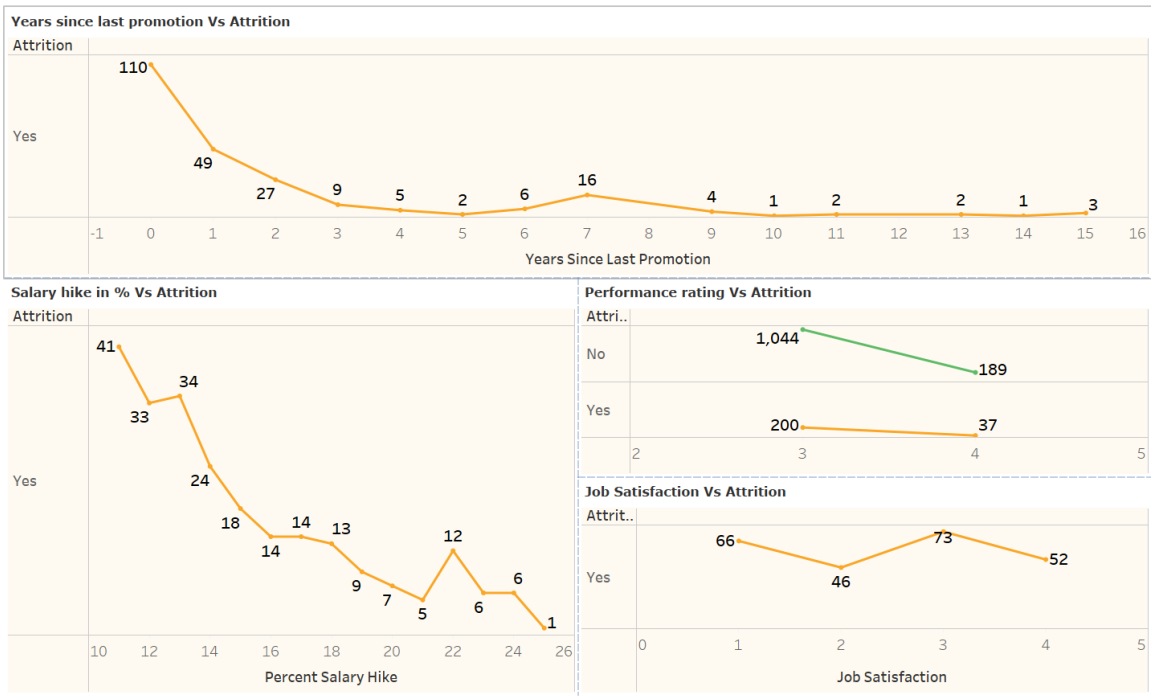


Figure 12 Attrition vs years since last promotion, salary hike, performance rating, and job satisfaction. (Tableau dashboard 3)

Dashboard 3 shows that 110 employees who recently got promoted quitted their jobs compared to those that spent more years working since their last promotion. While employees who received a low percent hike in salary had more attrition than those who a higher percentage increase in their salary. The falling trend highlights the fact that salary hike has a great influence on employee attrition.

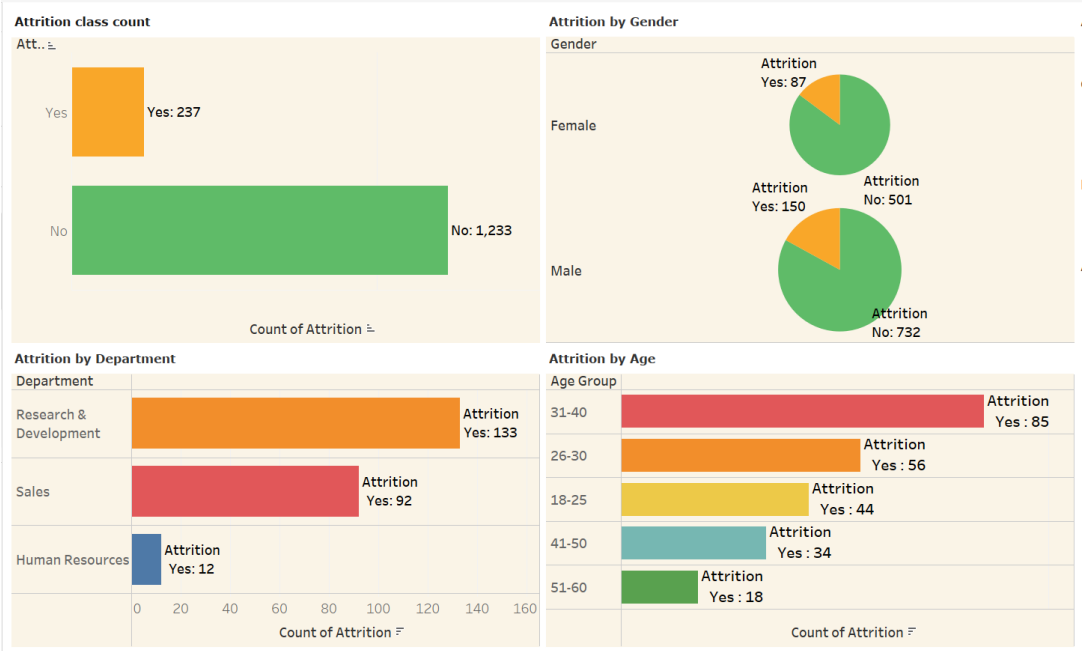


Figure 13 Attrition count vs gender, department, and age group. (Tableau dashboard 4)

In dashboard 4 shown in figure 13 above, we see the class distribution for all 1470 employees. Only 237 employees quitted their jobs. For the gender, the pie chart shows that the organisation has higher population of males than females, therefore, the males had more attrition than the females. Only 17% of the females and 20% of the number of males in the organization quitted the jobs.

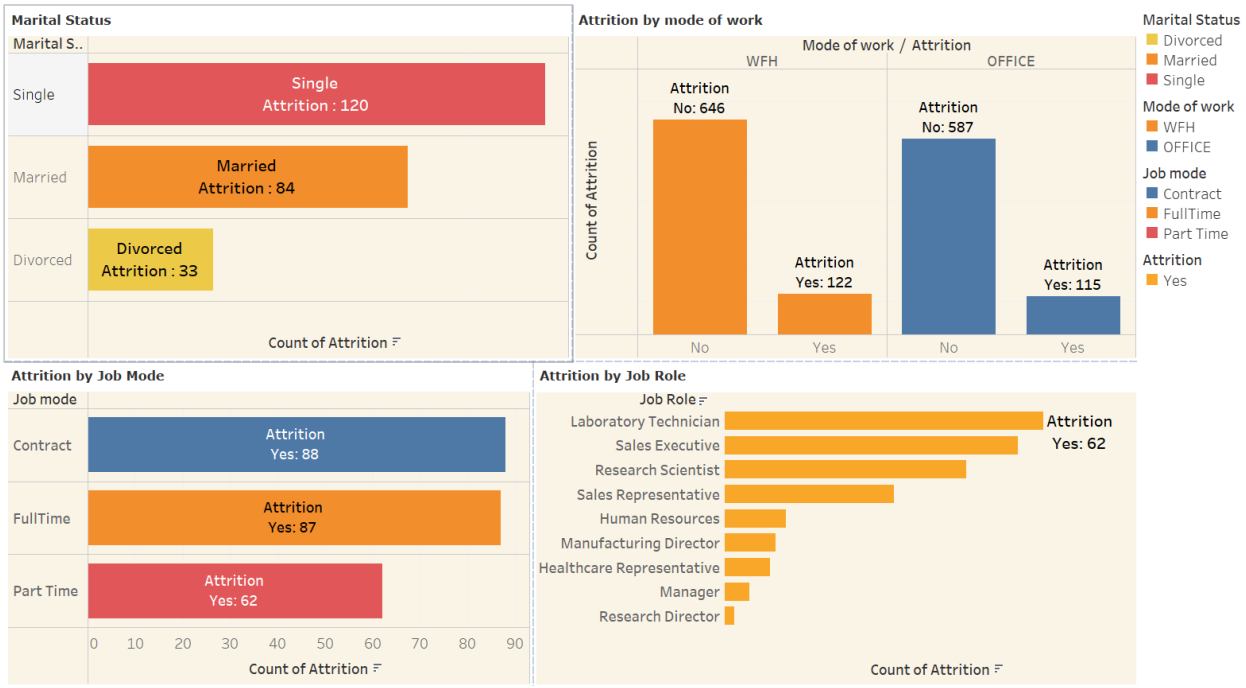


Figure 14 Attrition vs Marital status, mode of work, job mode, and jobe role. (Tableau dashboard 5)

In figure 14 above, we see the single ones have the highest attrition number, and those who work from home had higher number of attritions. For the job mode, those who are contract staffs and fulltime worker have higher attrition of 88 and 87 with fewer employees quitting as part-time. Employees with the job role of Laboratory technicians quitted more followed by the sales executive and research scientist respectively.

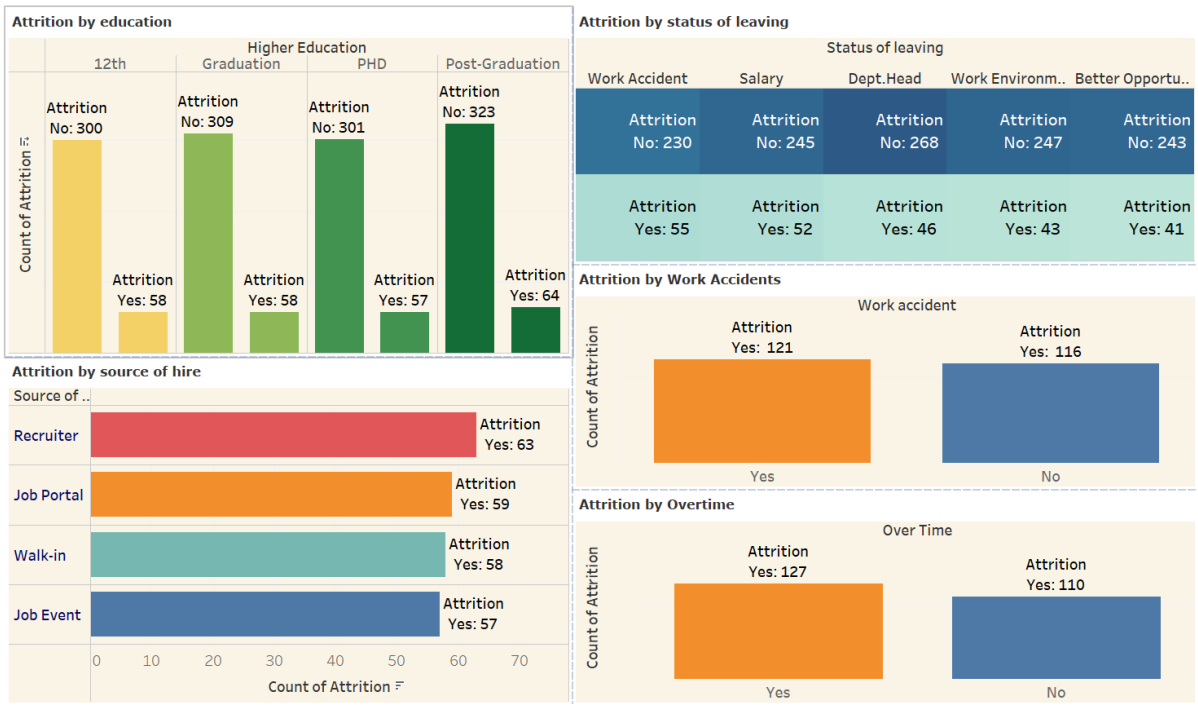


Figure 15 Attrition vs education level, status of leaving, work accidents, overtime, and source of hire. (Tableau dashboard 6)

From figure 15 above, we realized that employees with post-graduate education tends to attrite more compared to those with PHD, BSc and 12th grade. While the employees with work accident stated as reason for leaving quitted more which raises another sensitive question. Could it be that the organization doesn't observe safety protocols, or they don't compensate their staffs who were involved in accidents at work? This can also be said of the number of those that that did overtime at work. 127 of the employees that did overtime quitted their jobs, probably they didn't get the extra pay they expected.

These visual holds important information that gives insights concerning the various factors related to the attrition of employees.

3.3 Data Pre-processing

Data pre-processing is an essential step in Machine Learning because the quality of data and the useful information that can be extracted from it significantly influences our model's ability to learn. Thus, it is critical that we pre-process our data before feeding it into our model. The pre-processing methods that are employed in this experimental setup includes the following in order of execution:

- Data cleaning
- Data transformation
- Data encoding
- Data scaling or standardization

DATA CLEANING

In this experimental setup step, which involves modifying and removing data that is incorrect, duplicated, in wrong format, incomplete and irrelevant. I firstly checked for missing or null values because a dataset is not always accurate or organized. In machine learning, if the data is not thoroughly cleaned, the accuracy of the model stands on a weak and shaky foot. All the column in the dataset seems to be complete with no unnecessary duplications. Next step in this stage was to drop irrelevant columns based on common sense. Columns dropped are "Date of termination" and "Date hired" because they won't be necessary in constructing the machine learning model. This leaves us with 30 features.

DATA TRANSFORMATION

Recall the distributions of the continuous variables shown in figure 9, this is where we took care of those skewed distribution. Why did I care about these skewed data? If the data is skewed like those in figure 9, the model will be trained on a much larger number of moderately low values, and this will be less likely to successfully predict the values on the larger scale. The same is true with imbalanced categorical classes which implies that the model will be biased in its prediction right from the learning stage and this will violate the model's assumption or impair the interpretation of the feature importance. For this reason, I employed the use of the Box Cox transformation to help the data follow a normal distribution. The basic assumption of this transformation is that the data must be positive and continuous which is applicable in this case.

The original form of the Box Cox transformation is given by

$$x(\lambda) = \frac{(x^{\lambda}-1)}{\lambda} \; for \; \lambda \neq 0 \qquad \ldots\ldots\ldots\ldots\ldots\ldots 1$$

$$x(\lambda) = \ln(x) \ for \ \lambda = 0 \qquad \ldots\ldots\ldots\ldots..2$$

The extended form of the two parameter Box Cox transformation is given by

$$x(\lambda) = \frac{((x+\lambda_2)^{\lambda_1}-1)}{\lambda_1} \ for \ \lambda_1 \neq 0 \quad \ldots\ldots\ldots\ldots\ldots3$$

$$x(\lambda) = \ln(x + \lambda_2) \ for \ \lambda_1 = 0 \quad \ldots\ldots\ldots\ldots\ldots..4$$

Common Box-Cox transformations are as follows

Table 2

I	**Y'**
-2	$Y^{-2} = 1/Y^2$
-1	$Y^{-1} = 1/Y^1$
-0.5	$Y^{-0.5} = 1/\sqrt{Y}$
	Log(Y)
0.5	$Y^{0.5} = \sqrt{Y}$
1	$Y^1 = Y$
2	Y^2

Here are the results of the transformation for three features which are monthly income, total working years, and years with company respectively

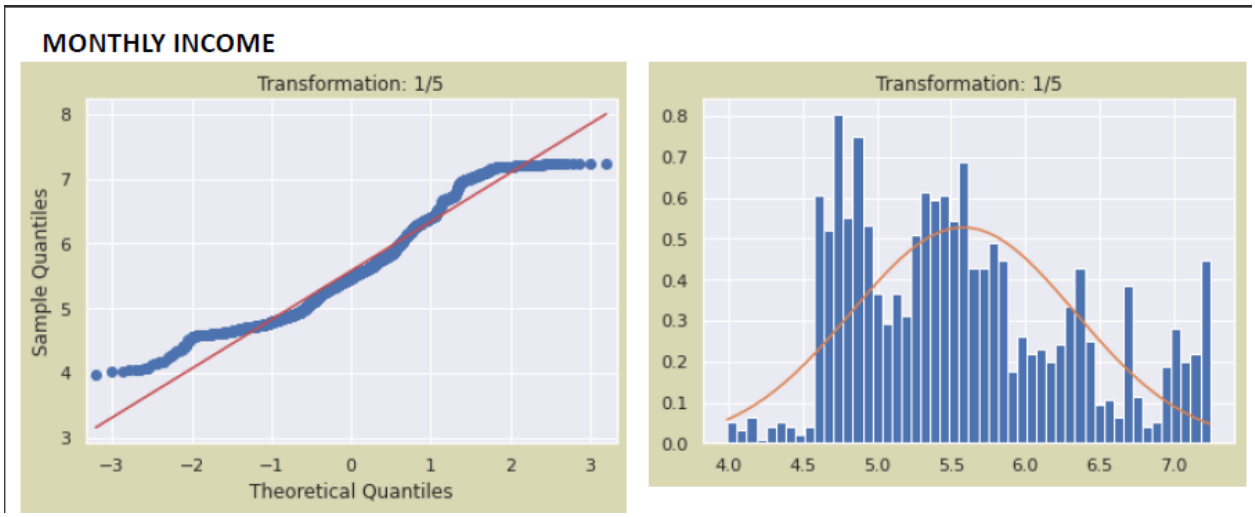


Figure 16 Box cox transformation of monthly income

The figure 16 shows the distribution after the transformation. I used the 1/5 transformation which was able to get the monthly income distribution as close to normal as possible. In figure 17, it shows the transformed distribution of the total working years and figure 18 shows the transformed data distribution for the years at company.

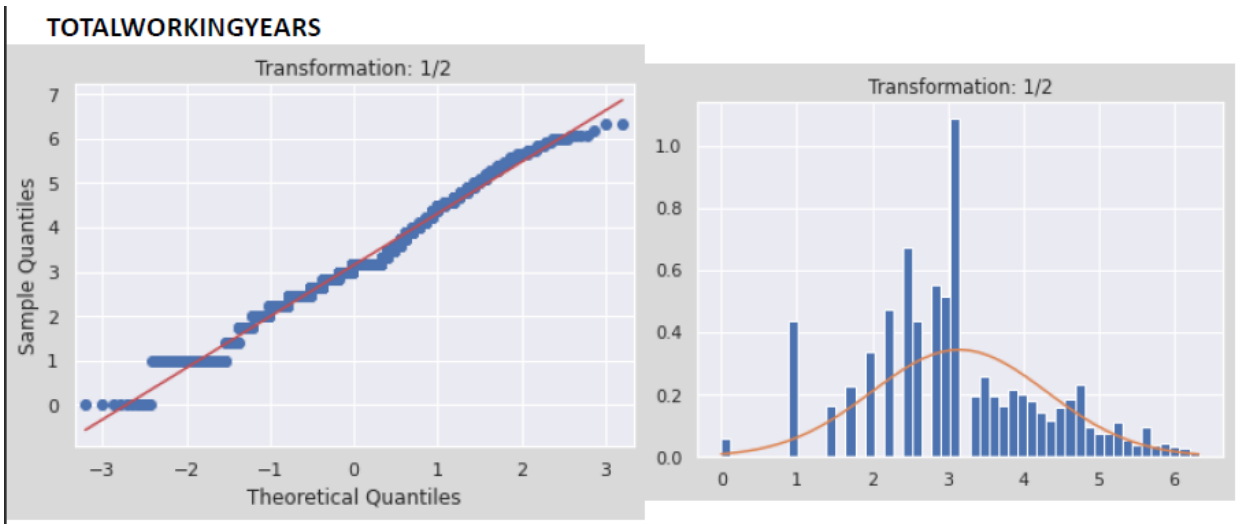


Figure 17 Box cox transformation of total working years

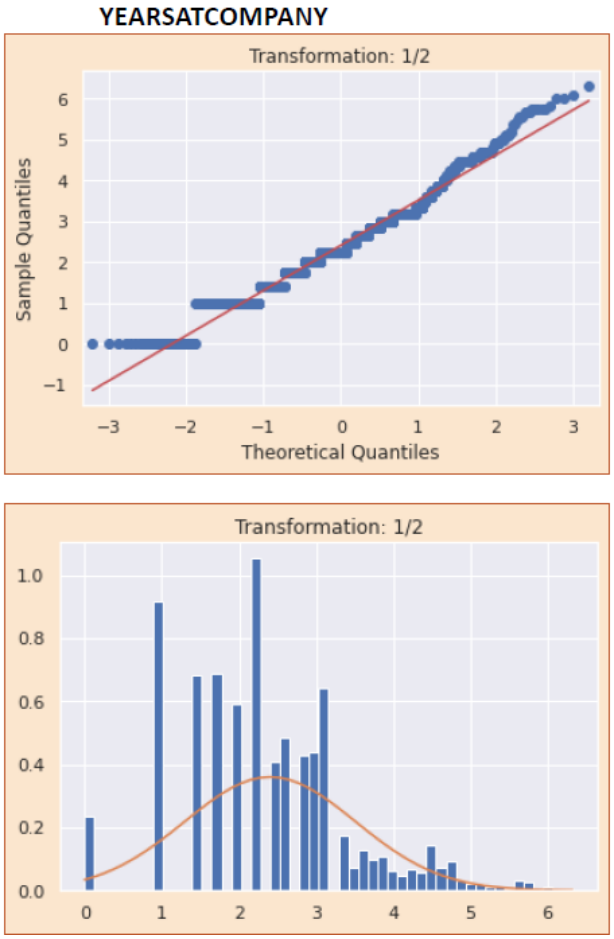


Figure 18 Box cox transformation of total years at company

DATA ENCODING

This is basically the process of converting the categorical data into digital signals for the secured transmission of data into the model. This is the only language the machine learning algorithm understands for it helps the algorithm perform better, additionally, it is a way of keeping the data safe from theft. The following categorical features were converted to digital signals:

"Attrition", "BusinessTravel", "Department", "Gender", "JobRole", "MaritalStatus", "OverTime", "Higher_Education", "Status_of_leaving", "Mode_of_work", "Work_accident", "Source_of_Hire", "Job_mode".

DATA SCALING OR STANDARDIZATION

The next step in the data pre-processing workflow is data standardization. In this context, standardization is used as a scaling technique to establish the mean and the standard deviation at 0 and 1 respectively. This presents the data in one common format.

Standardization:

$$z = \frac{x - \mu}{\sigma} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 5$$

With mean :

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots 6$$

And standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \quad \ldots\ldots\ldots\ldots\ldots 7$$

Where z is the z-score, x_i is the data point, μ is the mean, and σ is the standard deviation respectively.

3.4 Modelling

In this section I would be describing the model, its implementation, and the evaluation metrics I would be using in this project. The data is split as follows:

Table 3

Data set	Split fraction
Train set	60%
Validation set	20%
Test set	20%

Since this task involves predicting if an employee churned or not, it makes it a binary classification problem. It refers to a predictive modelling problem in a case where a class label, or output label, is predicted for a given example of input data and this makes it a supervised learning technique. This is represented in figure 19 below.

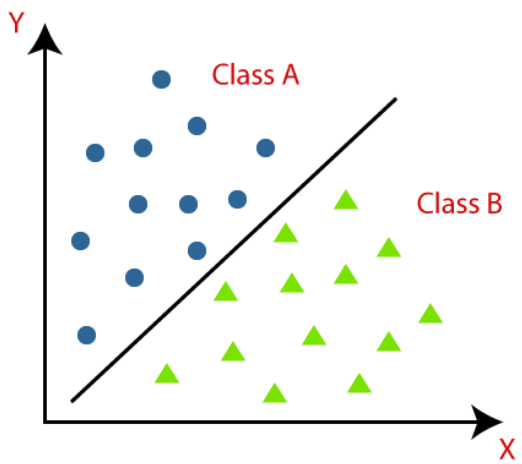


Figure 19 Binary classification

I employed two major techniques for this task for a proper defined comparison based on the aim and objective of this research project. They are the ensemble learning based technique and the machine learning based technique.

3.4.1 Ensemble learning based technique

The ultimate advantage of this technique is that it can make better predictions and achieve better performance than any single contributing model. Furthermore, it is also robust, reduces bias and variance to enhance the accuracy of the predictive model. Essentially, it is a process in which many different models are created to predict an outcome, either using many different modelling algorithms or using different training data sets or feature sets. It then aggregates each base model's prediction to produce a single final prediction for the unknown data set. This is depicted in figure 20 below.

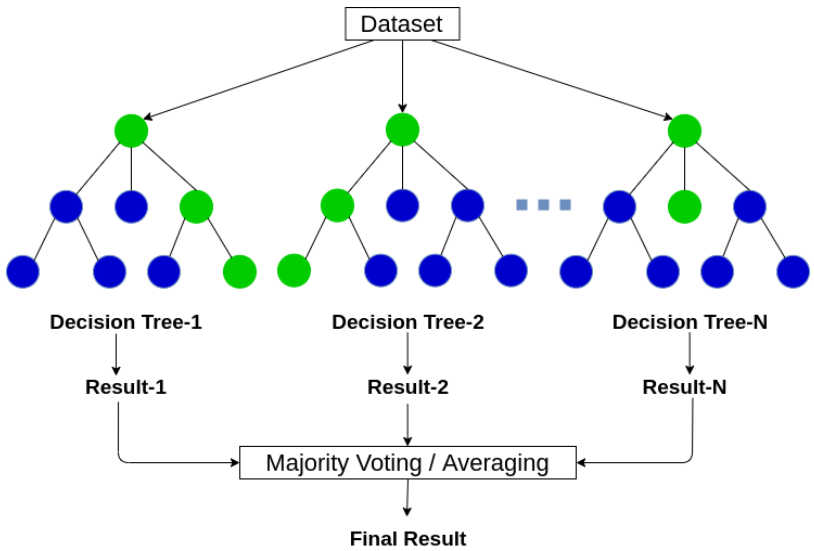


Figure 20 A typical ensemble learning based technique

The random forest algorithm is a classification algorithm made up of many decision trees. When building each individual tree, it employs bagging and feature randomness to create an uncorrelated forest of trees whose prediction as a synergy is more accurate than that of any individual tree. It runs efficiently on large data sets, provides an estimate of the predictive power of the features and their significance in the feature set, it provides an experimental method for identifying variable interrelations, and finally, it has methods for balancing error in an unbalanced class population. This makes it fitting for my classification task.

For the random forest classifier, I used a max depth of 2, 4, 8, 16, 32, and none while the number of estimators is given by:

$$2^I \qquad \text{where } I = 3 \text{ to } 10$$

I used a 5-fold and a 10-fold cross-validation and ran it on the feature set (will elaborate on this in the techniques section) and the best model was selected. Afterwards, I evaluated the models on the validation set and picked the best one. Finally, the best model was evaluated on the test set to gauge the model's ability to generalize to unseen data and to confirm its consistency. This is illustrated in figure 21 below.

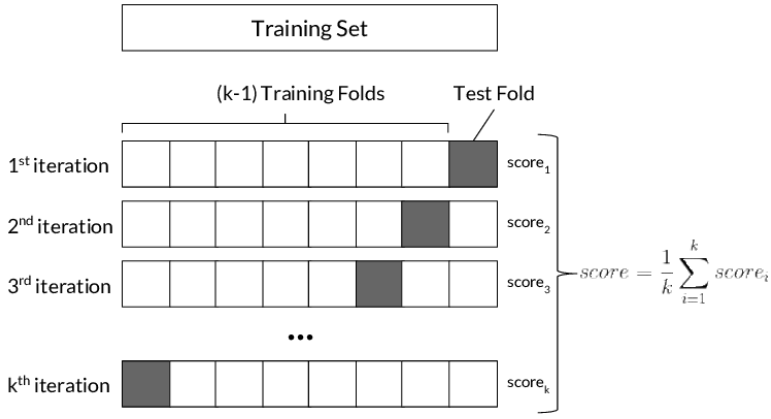


Figure 21 shows the kth-fold cross validation scheme

3.4.2 Machine learning based techniques: Logistics Regression

The most basic type of logistic regression is binary or binomial logistic regression, in which the target or dependent variable has only two possible values: 1 or 0. It enables us to model a relationship between multiple predictor variables and a binary/binomial target variable. In the case of logistic regression, the linear function is essentially used as an input to another function, such as g in the following relationship:

$$h_\theta(x)=g(\theta^T x) \text{ where } 0 \le h_\theta \le 1 \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots 8$$

Here, g is the logistic or sigmoid function which can be given as follows

$$g(z)=11+e^{-z} \text{ where } z=\theta^T x \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 9$$

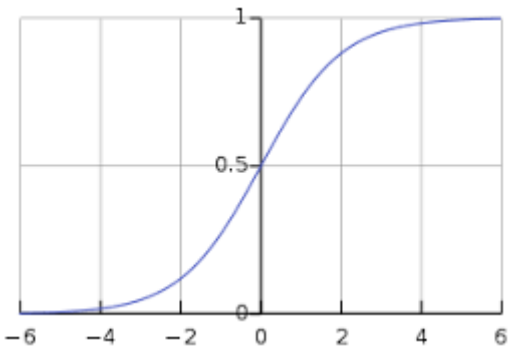


Figure 22 logistic regression

The sigmoid curve can be represented with the aid of a chart in figure. We can see the values of y-axis lie between 0 and 1 and crosses the x-axis at 0.5.

I also used a 5-fold cross validation method for this machine algorithm because I need it to determine the optimal parameters of the model which in this case is the "C" parameter. The goal is to estimate the generalization performance and stability of this learning procedure. The parameters are: Penalty = [I_1 , I_2], C = [0.01, 0.1, 1, 10, 100]

I used the logistic regression because it always has low variance, it is easy to apply and it does not take much computational duration.

3.4.3 Machine learning based techniques: Support vector machine

The support vector machine classifier, as it is widely known, is the most robust and accurate machine learning automated system among the other classification algorithms because it can efficiently perform a non-linear classification using the kernel trick, specifically mapping the inputs into high dimensional feature spaces.

3.4.4 Evaluation Metrics

Since it's a classification exercise, we would be focusing on the following evaluation metrics:

Metric	Description
Accuracy	This is the percentage of the correct prediction and its given by Accuracy = TP + TN / TP + TN + FP + FN
Precision	When the model predicts attrition, what percent of the time was it correct, that is, how many selected items are relevant. Its given by: Precision = TP / TP + FP
Recall	This describes how many relevant items are selected. It is the number of correct positive divided by the number of all samples that should have been identified correctly. It is given by: Recall = TP / TP + FN
F1- score	Is one of the most important evaluation metrics in machine learning, it elegantly sums up the predictive performance of a model by combining two otherwise competing metrics- precision and recall. It is a better metric when there are imbalance classes.
Latency	This tells us hold it took the computation to take place

4 Techniques

4.1 Feature Selection

Feature Selection is a method of reducing the number of input variables of your model by using only relevant data and removing noise in the data. It is the process of selecting relevant features for your machine learning model based on the type of problem you are attempting to solve inevitably. I employed this technique because this technique improves the machine learning procedure hence enhancing the predictive power of the machine learning output/result by selecting the most relevant features and eliminating the redundant ones.

For this task, I used the chi-square test which is specifically suitable for eliminating categorical features. Since the objective is to select the categorical variables that are highly reliant on the answer. It is a good option to use since it statistically tests the independence between variables. The formular for the chi square is:

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

Where: c = degree of freedom
O = observed values
E = Expected values

Since this method is a filter method, we calculated the chi-square between each feature and the target then proceeded to select the desired number of features with the best chi-square score or the lowest p-values as the case may be. The following categorical columns was assessed using this method to determine their predictive power: Attrition, business travel, department, gender, job role, marital status, overtime, higher education, status of leaving, mode of work, work accident, source of hire and job mode respectively.

Table 4 showing the chi-squared test result

features	p-values in descending order
Work accident	0.9588
Status of leaving	0.9585
Business travel	0.8857
Source of hire	0.8447
Higher education	0.8265
Mode of work	0.8056
Gender	0.7092
Department	0.5093
Job mode	0.1879
Job role	0.0907
Marital status	0.0007
Overtime	0.0000

As shown in the table 4 above, we see the p-values of each of these categorical features which describe the relationship these features have with the attrition. Our benchmark for screen the

irrelevant is the p-value of 0.5 and from the results job mode, job role, marital status, department, and overtime seems not to influence the predictive power of the algorithm.

4.2 Synthetic minority over-sampling technique (SMOTE)

In the exploratory data analysis section, we realised in figure 13 that out of 1,470 instances, we only had 237 persons quit their job and 1,233 never churned. This is a very sensitive situation of output class imbalance because it will definitely have a negative effect on the predictive performance of this classification algorithm. How do I mean, the algorithm will be biased thereby favouring the majority class which would yield a very high accuracy score but deep down we know that its result is particularly driven by the majority class which is not even our class of consideration. We are interested in predicting the minority class which is those that attrite. SMOTE is one effective solution to such issue in machine learning tasks. SMOTE works by picking instances in the feature space that are close together, drawing a line between them, and drawing a new sample at a point along that line.

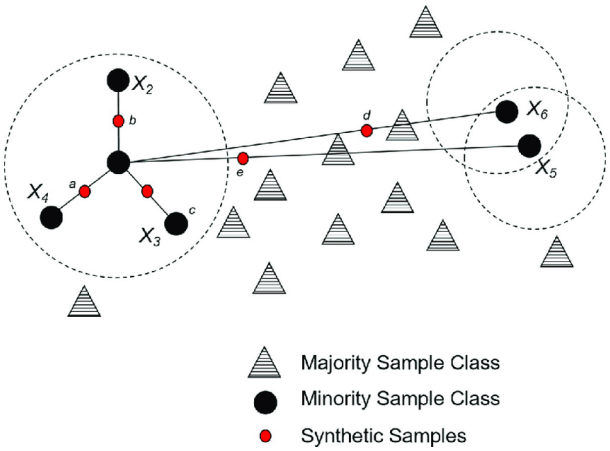


Figure 23 SMOTE operation

SMOTE begins by randomly selecting a minority class instance "a" and locating its k closest minority class neighbours. The synthetic instance is then generated by randomly selecting one of the k-nearest neighbours "b" and connecting a and b in the feature space to form a line segment. The synthetic instances are created by convexly combining the two selected instances, "a" and "b". The very good thing about this technique is that no information is lost or added to the dataset. This is illustrated in figure 23 above.

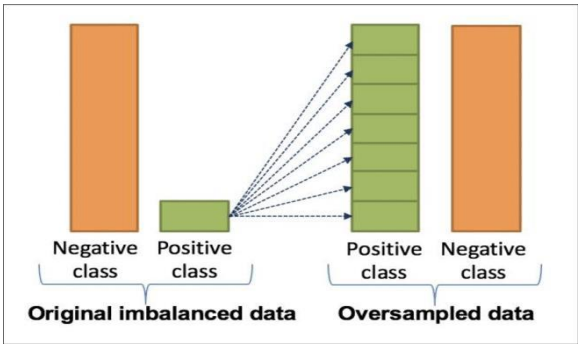


Figure 24 After smote class tends to be balanced

.

After executing the SMOTE technique, the result is as follows:

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

Here see that initially we had "0" (No attrition) = 735 while "1"(Yes attrition) = 147, but after applying SMOTE, both classes became the same. The minority class has increased hereby giving both classes a chance to be predicted with a more reasonable precision and accuracy without any fear of biased classification.

4.3 Feature Engineering

Now this is the primary technique this project is based on. The process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning is known as feature engineering. To make machine learning work well on new tasks, better features may need to be crafted and trained. It generally involves new feature creation, feature extraction as well as creating feature sets. For the Random Forest classifier which is an ensemble learning based algorithm that fits this task perfectly, I created 6 feature sets altogether from the dataset. As seen in the figure 25 below.

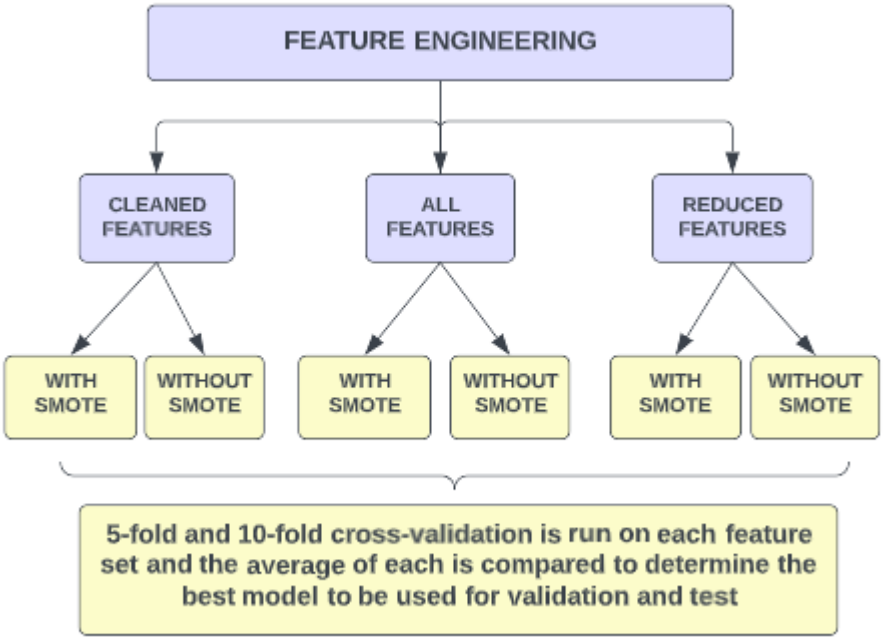


Figure 25 feature engineering work flow

The age column was replaced with Age_years while three extra features were created from existing ones after the box-cox transformation. I labelled them: monthlyIncomeTRF, TotalworkingYearsTRF, and finally YearsAtCompany respectively. After creating the three main feature sets cleaned feature set, reduced feature set and all feature set as seen in table 5, I went further to apply SMOTE to each of them to while keeping the previous feature set without

SMOTE applied. Just to mention, SMOTE was applied only on the train set but not on the validation set and the test set.

Table 5

Feature set description	Features
Cleaned features	'BusinessTravel', 'Department', 'DistanceFromHome', 'Gender', 'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'OverTime', 'PercentSalaryHike', 'PerformanceRating', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager', 'Higher_Education', 'Status_of_leaving', 'Mode_of_work', 'Leaves', 'Absenteeism', 'Work_accident', 'Source_of_Hire', 'Job_mode', 'Age_Years'
Reduced features	'BusinessTravel', 'DistanceFromHome', 'Gender', 'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'NumCompaniesWorked', 'PercentSalaryHike', 'StockOptionLevel', 'TrainingTimesLastYear', 'YearsSinceLastPromotion', 'YearsWithCurrManager', 'Higher_Education', 'Status_of_leaving', 'Mode_of_work', 'Leaves', 'Absenteeism', 'Work_accident', 'Source_of_Hire', 'Age_Years', 'MonthlyIncomeTRF', 'TotalWorkingYearsTRF', 'YearsAtCompanyTRF'
All feature set consisting of both cleaned and transformed together.	'BusinessTravel', 'Department', 'DistanceFromHome', 'Gender', 'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'OverTime', 'PercentSalaryHike', 'PerformanceRating', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager', 'Higher_Education', 'Status_of_leaving', 'Mode_of_work', 'Leaves', 'Absenteeism', 'Work_accident', 'Source_of_Hire', 'Job_mode', 'Age_Years', 'MonthlyIncomeTRF', 'TotalWorkingYearsTRF', 'YearsAtCompanyTRF'.

5 Result and Discussion

We would be comparing the 6 feature sets among each other based on just 3 of the performance metrics, recall, f1-score and latency. From table 6 observed that the cleaned feature set with the SMOTE applied both have the same accuracy of 86% at the training level for the 5-fold cross validation and for the 10-fold cross validation, the clean feature set with SMOTE applied gave the better accuracy. Result for the 5-fold and 10-fold cross validation is shown in table 6 and 7 below.

Table 6 Random Forest Classifier model train 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

Table 7 Random Forest Classifier model train result 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

Moving over to Validation stage where we evaluate the performance of the models on the validation set as started earlier as recorded in tables 8 and 9 respectively. For the 5-fold in terms of accuracy score, the SMOTE clean feature set and the all-feature set have the highest accuracy score of 85.7%. Compared to the f1-score, we see that SMOTE clean feature set produced the highest f1-score, ROC-AUC score and recall compared to the other feature set.

Table 8 Random Forest classifier model results on validation set 5-fold cv

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

Looking at table 9 below representing the 10-fold cross validation, we notice that SMOTE clean feature set also has the highest accuracy score, and ROC-AUC score. When comparing the latency, the reduced feature set without SMOTE tends to be the fastest in computational speed. This agrees to the fact that a reduced feature can help in reducing the time spent in computing which is a priority in many areas of data analysis.

Table 9 Random forest classifier model results on validation set 10-fold cv

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

Table 10 Random forest classifier model results on test set 5-fold cv

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

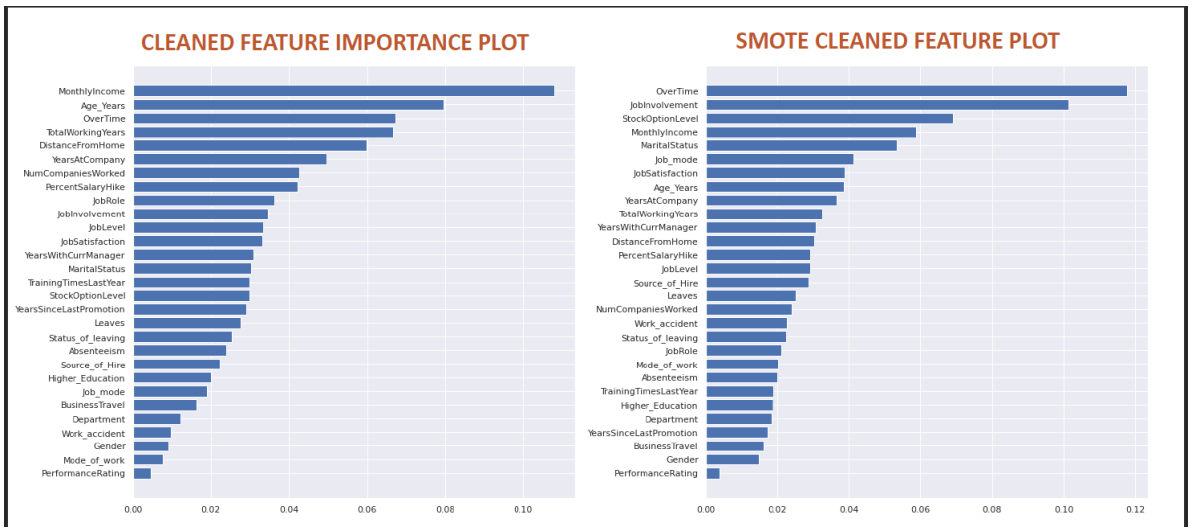
Did you notice how the overall performance improved as we go through each evaluation stage? Yes, this is proven that ensemble learning methods is efficient although requires a lot of work and commitment. Smote clean feature set have a higher accuracy score, precision, recall, f1-score, ROC-AUC as well. Its accuracy increased from 85% at the training stage to 87% at the test stage. Precision also increased from 60% to 65%. This is worthwhile because the whole research exercise is not about getting a correct result or not. From these outcomes I could say to an extent that these combined techniques really had an impact on the performance of the model with excellent significant result.

Table 11 Random Forest classifier model results on test set10-fold cv

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

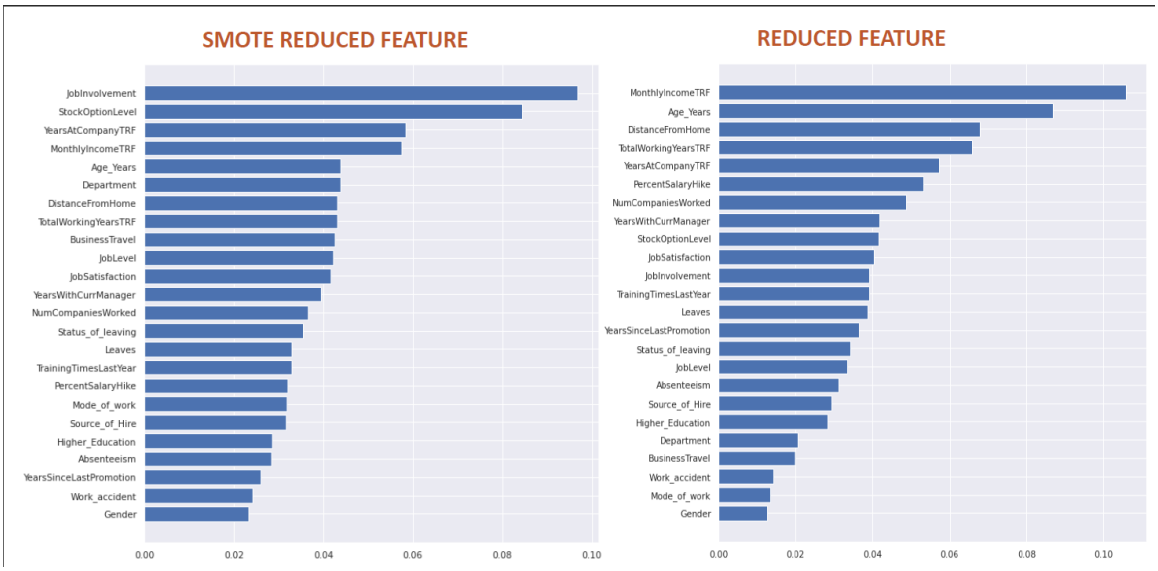
Same cannot be said about the 10-fold cross validation results. The results show that the values for the performance metric scores reduced or had no significant change.

Figure 26 shows the cleaned feature importance plot with and without SMOTE



Using the Selectkbest for feature importance plot which is one of the good things about random forest classifier. Being able to determine which feature had greatest impact on the class prediction. from figure 26 above, we see that features like monthly income, age, overtime, job involvement, total working years, percentage salary hike obviously have greater impact in predicting if an employee would leave or not.

Figure 27 shows the reduced feature importance plot with and without SMOTE



In the plot comparing SMOTE reduced feature set in figure 27, we also see the reoccurence of some features that confirms its relevance to predicting employee attrition. We see job involvment, monthly income, age, salary bike, distance from home, years at company and so on.

This way we could determine with a level of certainty that this feature that show up more frequently on this feature importance plots have greatest impact on the model's predictive powers compared to the others. This lays a groundwork for further study. I will talk more about it in future work and accomplishments section.

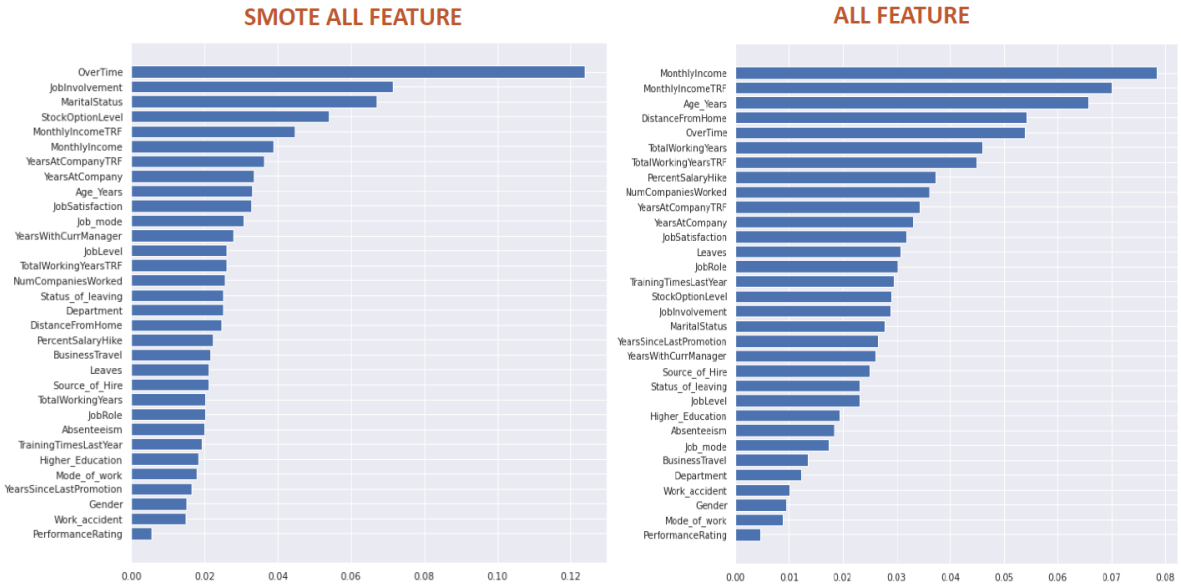


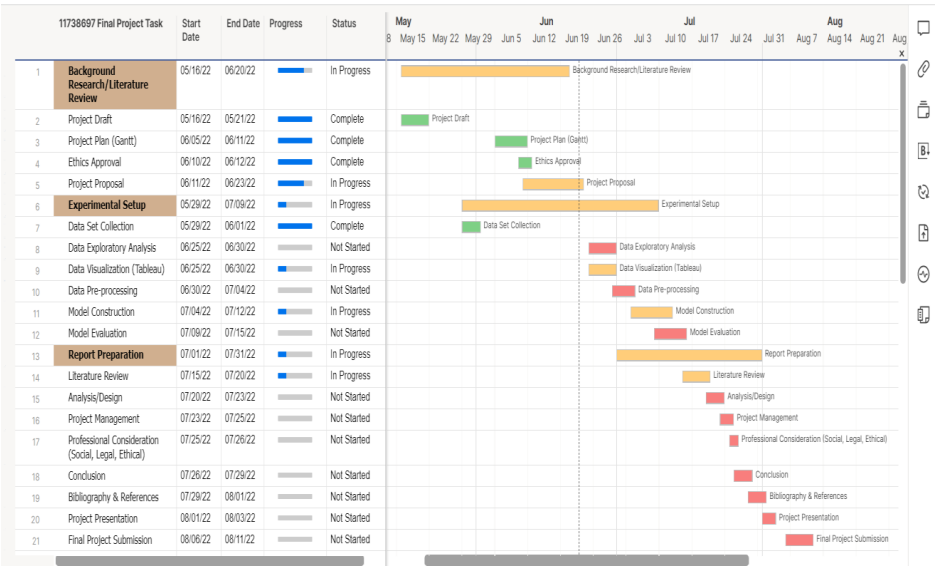
Figure 28 shows the reduced feature importance plot with and without SMOTE

6 Project Management

6.1 PROJECT SCHEDULE

In order for me to be successful in achieving my aims and objectives which involved executing and achieving the result for the proposed experimental design, a schedule on a Gantt chart was designed from the onset. It highlighted my plan about execution of specific task which kicked off on 15th of May 2022.

Figure 29 Gantt chart show project plan and schedule



	11738697 Final Project Task	Start Date	End Date	Progress	Status
1	**Background Research/Literature Review**	05/16/22	06/20/22		In Progress
2	Project Draft	05/16/22	05/21/22		Complete
3	Project Plan (Gantt)	06/05/22	06/11/22		Complete
4	Ethics Approval	06/10/22	06/12/22		Complete
5	Project Proposal	06/11/22	06/23/22		In Progress
6	**Experimental Setup**	05/29/22	07/09/22		In Progress
7	Data Set Collection	05/29/22	06/01/22		Complete
8	Data Exploratory Analysis	06/25/22	06/30/22		Not Started
9	Data Visualization (Tableau)	06/25/22	06/30/22		In Progress
10	Data Pre-processing	06/30/22	07/04/22		Not Started
11	Model Construction	07/04/22	07/12/22		In Progress
12	Model Evaluation	07/09/22	07/15/22		Not Started
13	**Report Preparation**	07/01/22	07/31/22		In Progress
14	Literature Review	07/15/22	07/20/22		In Progress
15	Analysis/Design	07/20/22	07/23/22		Not Started
16	Project Management	07/23/22	07/25/22		Not Started
17	Professional Consideration (Social, Legal, Ethical)	07/25/22	07/26/22		Not Started
18	Conclusion	07/26/22	07/29/22		Not Started
19	Bibliography & References	07/29/22	08/01/22		Not Started
20	Project Presentation	08/01/22	08/03/22		Not Started
21	Final Project Submission	08/06/22	08/11/22		Not Started

The Gantt chart in figure 29 shows the task, the progress and the time allocated for the tasks. The ethics certificate, project proposal and the presentation slides are included in the appendix of the project report for reference purposes.

6.2 RISK MANAGEMENT

I must say this aspect of the project got me at the edge of my sit because I feared losing my data, my codes and one time my personal computer crashed. However, in foresight, precautionary steps were taken to curtail any form of risk that could affected the progress of this project and completion of this project. Regular and multiple back up was done to avoid losing data. That includes using a code editor that is linked to my google account where all my codes are safe and unaltered in anyway. Google colab services was use through out for the coding and Tableau visual app was used to create the visualizations for the exploratory data analysis. Both apps gave the opportunity to back up regularly to ensure my work was safe.

6.3 QUALITY MANAGEMENT

Contact was maintained with the supervisor, both virtually and in person, to share progress and monitor the quality of the project. More to this effect is included in the meeting records in appendix C.

6.4 Social, Legal, Ethical and professional considerations

Prior to the start of this project, issues concerning the legal, ethical and social implications was put into account, owing to the fact that as data scientist, we often love to play around with data depending on the task at hand. In some case this data might include sensitive information that should be kept confidential at all costs. My data set, as mentioned from the onset, does not contain any personal information like name, address, health records or financial details. It was gotten from a public data repository and was stated in the ethics form which was approved with low threat. Th ethics certificate is also attached this document in appendix A.

7 Conclusion

7.1 Achievements

During this project till its near completion, I must say the objective and aim was met despite the challenges due to trying out a combination of advanced machine learning techniques. I was able to show by my work that advanced feature engineering combined with SMOTE and feature selection methods can really have an impact on a machine learning model. I must say it was quite tough a little bit because I feared I was not going to get reasonable results that could back up my approach to a reasonable extent.

Moreover, it can be said that the aim and objective stated at the onset of this project journey was satisfied, and it has paved way for more critical thinking on how to approach human resource related issues using machine learning techniques that best fit the challenge at hand.

Conclusively, I can confidently say that the ensemble learning based technique which allows the comparison of multiple models with the best being chosen after careful evaluation is an effective method and the analysis result attest to it.

7.2 Future work

Despite the dept of analysis carried out in this project there is still more work to be done. The field of data analysis keeps evolving and this give all analyst and scientist a chance to efficiently solve difficult problems facing digital world, with speed and accuracy. As future work, I intend to propose way to further improve the predictive performance of this classification models considering lesser but influential features. Propose ways to achieve more accurate prediction with shorter computational runtime.

Propose ways to tune hyper-parameters to optimize model performance, trying out other combination of techniques to achieve same aim with higher accuracy, lesser time.

8 Student reflections

Throughout progress of this project, I pushed myself beyond my limit. Fear of the possible outcome, distractions and challenges tried to weigh me down but I realized that my strength and capabilities are more woke in such times and that pushed me into see this project to the end at my best capabilities. My analytics skills have improved tremendously as well as my coding skill. Understanding how to approach a problem, where to look for solutions, proper organization, management and timing. I must say I am better now and I will not stop until I achieve something great that no one else have thought about let alone do. As much as it was engaging, tough and promising at the same time, I must say I have more to be thankful about than complain because I have gained knowledge which in turn has given power to explore and achieve greater heights.

Bibliography and References

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

R. Punnoose and P. Ajit. (2016). ''*Prediction of employee turnover in organizations using machine learning algorithms*'' Int. J. Adv. Res. Artif. Intell. vol. 5, no. 9, p. 5. doi: 10.14569/IJARAI.2016.050904.

Kerim Koc , Omer ¨ Ekmekcioglu , Asli Pelin Gurgun. (2021). "Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers". Page 9. https://doi.org/10.1016/j.autcon.2021.103896

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

Shuai Ye, Guoen Xia, Xianquan Zhang, Xiang Qing Li. School of Computer Science and Engineering, Guangxi Normal University, Guilin. (2021). "Application of Hybrid Sampling Method in the Prediction of Telecom Customer Churn". Page 2 &3. https://dl.acm.org/doi/pdf/10.1145/3508546.3508597

M. Subhashini and R. Gopinath, Employee Attrition Prediction in Industry Using Machine Learning Techniques, International Journal of Advanced Research in Engineering and Technology, 11(12), 2020, pp. 3329-3341. https://iaeme.com/Home/issue/IJARET?Volume=11&Issue=12

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

Rupa Chatterjee Das Anjana Devi S.C. (2020). "*Conceptualizing the Importance of HR Analytics in Attrition Reduction*". International Research Journal on Advanced Science Hub, 2(Special Issue ICAMET 10S), 40-48. doi: 10.47392/irjash.2020.197

Sujeet N.Mishra, Dev Raghvendra Lama, Yogesh Pal. (2016). "*Human Resource Predictive Analytics (HRPA) for HR Management in Organizations* ". International Journal of Scientific & Technology Reseach, Vol 5, issue 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

Fattahi, S., Othman, Z., and Othman, Z. A. 2015. NEW APPROACH WITH ENSEMBLE METHOD TO ADDRESS CLASS IMBALANCE PROBLEM.Journal of Theoretical & Applied Information Technology. 72,1.

Peng, L., Zhang, H., Yang, B., and Chen, Y. 2014." *A new approach for imbalanced data classification based on data gravitation*". Information Sciences. 288, 347-373.

Sun, Y., Wong, A. K., and Kamel, M. S. 2009. *Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence*. 23,04, 687-719. https://us04web.zoom.us/j/9931191963?pwd=lGr0j6eXlreeDRpzMcSErz0ObXviuA.1

Tanmay Prakash Salunkhe. 2018. "*Improving Employee Retention by Predicting Employee Attrition using Machine Learning Techniques*". https://esource.dbs.ie/bitstream/handle/10788/3497/msc_salunkhe_t_p_2018.pdf?sequence=1&isAllowed=y

A. Amin, F. R. I. A. C. K. a. S. A., 2015. *A Comparison of Two Oversampling Techniques (SMOTE vs MTDF) for Handling Class Imbalance Problem*: A Case Study of Customer Churn Prediction. Cham: Springer International Publishing, p. 215–225.

Kashyap. B, Kriti. S. 2018. "*Comparative study of the machine learning techniques for predicting the employee attrition*". Volume 5, Issue 3.E-ISSN 2348-1269, P- ISSN 2349-5138 . https://ijrar.org/papers/IJRAR1903202.pdf

R. Shiva Shankar, V. Priyadarshini, P. Neelima, CH. Raminaidu, 2021. "*Analyzing Attrition and Performance of an Employee using Machine Learning Techniques*", *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp.1601-1608.

Mas Rahayu Mohamad, Fariza Hanum Nasaruddin, Suraya Hamid, Sarah Bukhari, Mohamad Taha Ijab, 2021 "*Predicting Employees' Turnover in IT Industry using Classification Method with Feature Selection*", 2021 International Conference on Computer Science and Engineering (IC2SE), vol.1, pp.1-7.

Fatma Ozdemir, Mustafa Coskun, Cengiz Gezer, and V. Cagri Gungor. 2020. "Assessing Employee Attrition Using Classifications Algorithms". In Proceedings of the 2020 the 4th International Conference on Information System and Data Mining (ICISDM 2020). Association for Computing Machinery, New York, NY, USA, 118–122. https://doi.org/10.1145/3404663.3404681

Derek Jedamski. 2020. "*Applied Machine Learning Feature Engineering*". https://www.linkedin.com/learning/applied-machine-learning-feature-engineering

Jason Brownlee. 2020. "*How to choose a Feature Selection Method for Machine Learning*" . Machine learning mastery https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Appendix A – Project Proposal

Computing Individual Research Project

(7151CEM)

Project Proposal

Data Science & Computational Intelligence (MSc)

Project Title

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

Author
UCHECHUKWU FREDRICK OKONKWO
(SID): 11738697

Supervisor
VERNON SASSMAN

Faculty of Engineering, Environment and Computing

School of Computing, Electronics and Mathematics

Coventry University

Submission date

24th June 2022

Contents

1 Problem Statement and Research Question Error! Bookmark not defined.
1.1 Problem Statement ..... Error! Bookmark not defined.
1.2 Project aim/objectives ..... Error! Bookmark not defined.
1.3 Research question 3
2 Intended user or groups of users and their requirements 3
3 Systems requirements and project deliverables 3
4 Primary Research Plan 4
4.1 Background Research/Literature Review 4
4.2 Data Collection 4
4.3 Exploratory Data Analysis 4
4.4 Data Pre-processing 4
4.5 Model Construction 5
4.6 Model Evaluation 5
4.7 Final Report Preparation 5
5 Initial Literature Review 5
6 Bibliography 7
7 Appendix 8

Ethics Progress

☒	I submitted my ethics application, and my application has been approved. I include my ethics certificate in the appendix as evidence
☐	I submitted my ethics application, and my application is currently under review.
☐	I have not submitted my ethics application.

1 PROBLEM STATEMENT AND RESEARCH QUESTION

1.1 PROBLEM STATEMENT

Attrition prediction is essentially predicting which clients are most likely to quit working for an employer based on some relevant factors like Age, marital status, general performance, job satisfaction and so on. From the company's point of view, employee attrition is serious in HR analytics. It is necessary to gain this information because recruiting employees is often arduous and costlier than retaining old ones. It also necessary because when trained, technical and resourceful employees leave a place of employment for another job which offers better opportunities, this often results in financial loss and low morale of remaining employees. 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. Therefore, companies make use of this information to see how they can prevent the employee from leaving by improving the employee retention.

1.2 PROJECT AIM AND OBJECTIVE

In this project, I will be using various machine learning classification algorithms and feature engineering to predict if an employee would quit his/her job or not eventually and finally evaluate and select the best model. Additional aim is to identify the most influential factors surrounding the attrition of employees and its possible variations.

1.3 RESEARCH QUESTION

What factors or features have more influence on the attrition rate of employees? Does feature engineering or feature selection enhance the model's accuracy, precision, or overall performance in anyway irrespective of the algorithm used?

2 INTENDED GROUP OF USERS AND THEIR REQUIRMENTS

Those who stand to benefit maximally from this project are the human resources managers as well as the project managers. Applying this comparative machine learning predictive analysis on the employee's data gives the HR managers control over the attrition rate saving the company a lot of cost and resources. The general idea is to improve retention rate. However, to achieve this, careful decision must be made on which employees are valuable and profitable to the organization so they can be retained.

3 SYSTEMS REQUIRMENTS AND PROJECT DELIVERABLES

3.1 Project draft preparation & Initial background research
Deliverables: Project proposal and project plan.
Aim: Research, study and review most recent related research work to identify the setbacks. Get appropriate data set and carefully examine its credibility.

3.2 Implementation and Model design
Deliverable: Circumscribe the experimental process in developing the model. Pre-processing (feature engineering), modelling, training, and validation of model.
Aim: Define evaluation of performance metrics to decide best model performance. Parameter tuning

3.3 Model testing
Deliverable: Test results, metrics comparison and evaluation.
Aim: Evaluate model, feature selection (backward elimination), saving best model for deployment or future use.

3.4 Result and discussion
Deliverables: Project final documentation, experiment result presentation, and conclusions

4 PRIMARY RESEARCH PLAN

4.1 BACKGROUND RESEARCH/LITERATURE REVIEW

This task involves intensive research to find answers to the research question with the aim of finding most recent and relevant information from the primary research paper sources such as google scholar, association for computing machinery ACM digital libraries, and Institute of electrical and electronics engineers-IEEE Xplore digital libraries. At least 4 papers from these sources related to employee attrition prediction will be selected for review.

4.2 DATA COLLECTION

The data set to be used for this research project will be gotten from Kaggle online data repository which is public (dataset information can be found in appendix). This data set is a human resource data based on IBM attrition which contains 32 features and 1470 rows. The data set features consist of 17 integers, 11 strings, 3 Booleans and 1 other data type.

4.3 EXPLORATORY DATA ANALYSIS AND VISUALIZATIONS

This is basically analysing and investigating the data set and summarizing their main characteristics to discover patterns, anomalies, or verify assumptions. Often employing data visualization methods like correlations, heatmaps, graphs and plots. I will be using the Tableau software for part of the exploratory data analysis/visualization.

4.4 DATA PRE-PROCESSING

This one of the most important and sensitive stage of this project which involves cleaning, transformation (scaling/normalization, splitting & grouping features, encoding categorical features) of the data set. This describes in part the feature engineering process. The data set has an imbalance target variable which will influence the model's prediction result; hence, I will employ the use of the Synthetic Minority Over-sampling technique (SMOTE) which is a well-known technique for tackling the issue of data imbalance.

4.5 MODEL CONSTRUCTION, IMPLEMENTATION AND EXPERIMENT

I will be using the Gradient boost classifier, Logistic regression classifier, K-Nearest Neighbour, Random Forest, and the support vector machine classification algorithms. This task covers the data set splitting, model training, validation, model testing, and feature selection (backward elimination, correlation coefficients, and chi-square). The models will be fitted on the data set with raw features, cleaned/transformed features, all features, and reduced features. The python programming language will be used for all this tasks because of its versatility and flexibility.

4.6 MODEL EVALUATION

The major performance metrics to be used for the evaluation of the models are confusion matrix, ROC, AUC, accuracy, precision, recall, model latency and the f-score respectively.

4.7 FINAL REPORT PREPARATION

This will cover the entire research and experimental process from the literature review to the results. It will be well documented according to the academic research report standard.

5 INITIAL LITERATURE REVIEW

Employee attrition or voluntary turnover is a significant concern for most organization with a functioning HR department as it affects not just their efficiency or work sustainability but also their long-term development tactics. For this reason, many researchers have employed various techniques into people analytics to comprehend and explain employee attrition so they could recommend solutions to help human resource managers to reduce attrition by improving the way they recruit and retain resourceful employees. The task is predictive analysis which is a branch of advanced analytics that makes predictions about future outcomes based on historical data combined with statistical modelling, feature engineering techniques and machine learning as a synergy. Organizations employ predictive analytics to find patterns in the data to identify problems, opportunities, and underlying circumstances, hence, proffer specific solutions centred on the analysis outcome.

Research reports show several employee attrition or voluntary turnover predictive models as well as various approach to techniques with the major aim to enhance the model performance metrics. For this binary class prediction task, researchers have used logistic regression, decision tree, support vector machines, K-nearest neighbour, naïve bayes, XGboost and random forest classifier with a varying combination of dimension reduction (principal component analysis), feature selection, model optimization, esembling methods and SMOTE for data imbalance problems [2], [3], [4].

Nesrine, Jihen and Ricardo used the following classifiers: Decision Tree, Logistic Regression and Support Vector Machine (as machine learning models), Random Forest, XGBoost and Vote Classifier (as ensemble learning models) and three deep learning models (DNN, LSTM and CNN). A grid search algorithm was performed for each classifier over tuning hyperparameters. Then, the different models were trained using their best configuration on the training dataset. They evaluated their results, and it showed the outcome was improved after two feature selection techniques, Recursive Feature Elimination (a wrapper method) and SelectKBest (a filter method), was applied in the analysis [1].

In my project I will be combining the feature engineering technique which is basically the process of transforming the raw data into features that better represent the underlying signal to be fed to a machine learning model, SMOTE to balance class distribution by randomly increasing minority class examples which creates larger and less specific decision boundaries that increase the generalization capabilities of classifiers, majority voting ensemble to combine the prediction from multiple other models to improve product performance, ideally achieving better performance than any single model used in the ensemble. I will fit the model to the data set with the raw features, the data set with cleaned and transformed features, a combination of both and finally a data set with selected feature.

The aim of combining these techniques together is to conclude if they will have an impact on the classification prediction accuracy either positively or negatively and see what other things we can learn about the individual feature's influence on the employee attrition.

6 BIBLIOGRAPHY

[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] R. Punnoose and P. Ajit. (2016). ‘‘Prediction of employee turnover in organizations using machine learning algorithms,’’ Int. J. Adv. Res. Artif. Intell. vol. 5, no. 9, p. 5. doi: 10.14569/IJARAI.2016.050904.

[3] Kerim Koc , Omer ¨ Ekmekcioglu , Asli Pelin Gurgun. (2021). “Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers”. Page 9. https://doi.org/10.1016/j.autcon.2021.103896

[4] 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

[5] Shuai Ye, Guoen Xia, Xianquan Zhang, Xiang Qing Li. School of Computer Science and Engineering, Guangxi Normal University, Guilin. (2021). “Application of Hybrid Sampling Method in the Prediction of Telecom Customer Churn”. Page 2 &3. https://dl.acm.org/doi/pdf/10.1145/3508546.3508597

[6] M. Subhashini and R. Gopinath, Employee Attrition Prediction in Industry Using Machine Learning Techniques, International Journal of Advanced Research in Engineering and Technology, 11(12), 2020, pp. 3329-3341. https://iaeme.com/Home/issue/IJARET?Volume=11&Issue=12

[7] 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

7.2 Certificate of Ethical Approval

Coventry University

Certificate of Ethical Approval

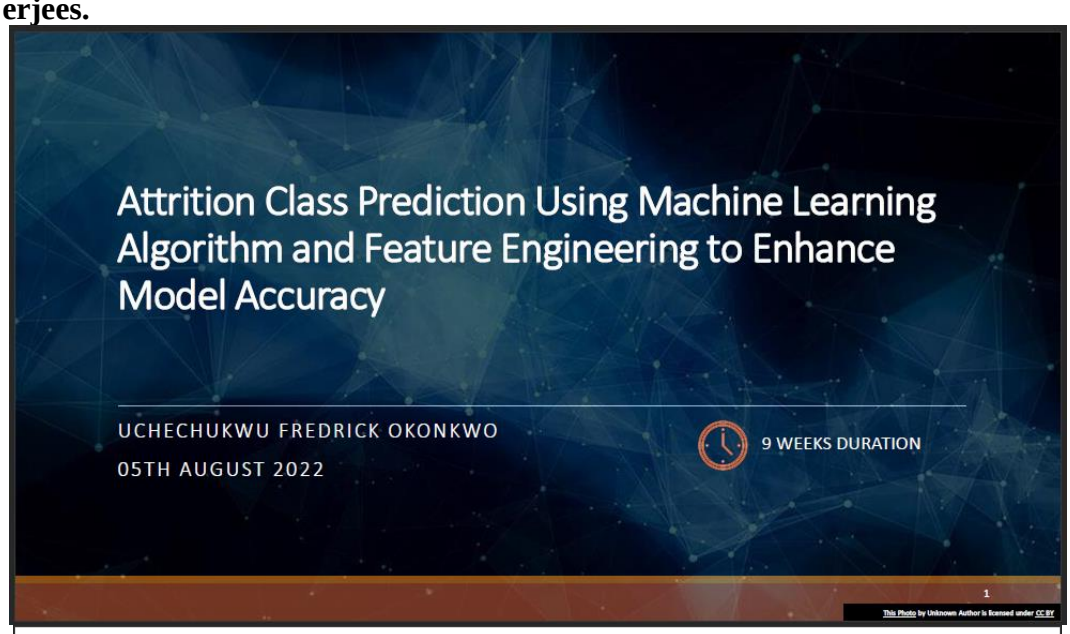
Applicant:	Uchechukwu Okonkwo
Project Title:	Attrition class Prediction using machine learning and feature engineering to enhance model accuracy

This is to certify that the above named applicant has completed the Coventry University Ethical Approval process and their project has been confirmed and approved as Low Risk

Date of approval:	12 Jun 2022
Project Reference Number:	P137917

Appendix B – Project Presentation Slides

Presented on 11/08/2022 at 10:05am – 10:35am. In the presence of Dr Vernon Sassman and Zina Jerjees.



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

4

OBJECTIVE

5

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?

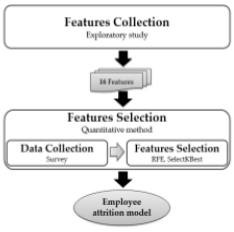
6

RELATED WORK

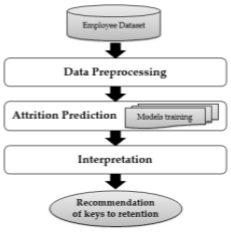
7

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**

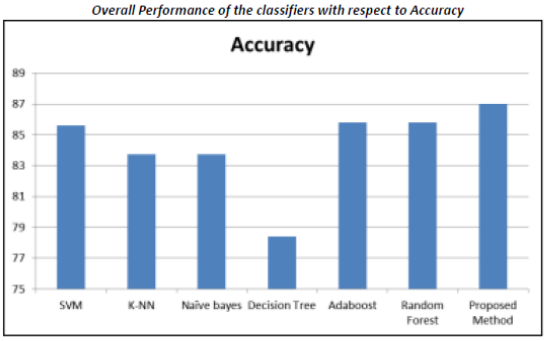
8

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

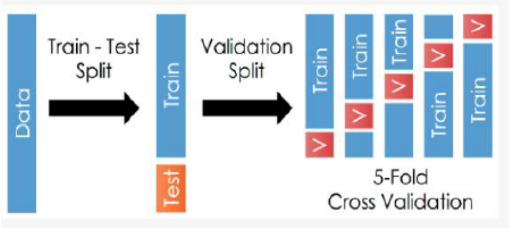
Performance Measure Metrics	SVM	K-NN	Naïve Bayes	Decision Tree	Adaboost	Random Forest
Accuracy	85.6%	83.74%	83.74%	78.4%	85.8%	85.8%
MSE	0.144	0.1626	0.16	0.216	0.14	0.142

Overall Performance of the classifiers with respect to Accuracy



9

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



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

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

10

DATA SET, EDA & VISUALIZATION

11

```
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
```

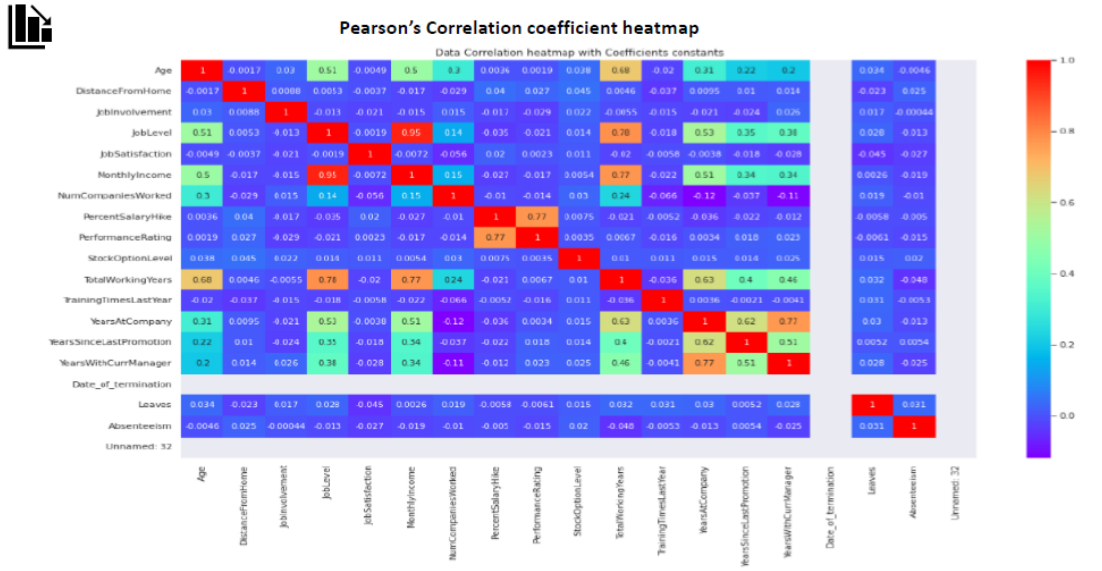
HR Attrition data based on IBM attrition.

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

It consist of 32 columns with 1470 instances.

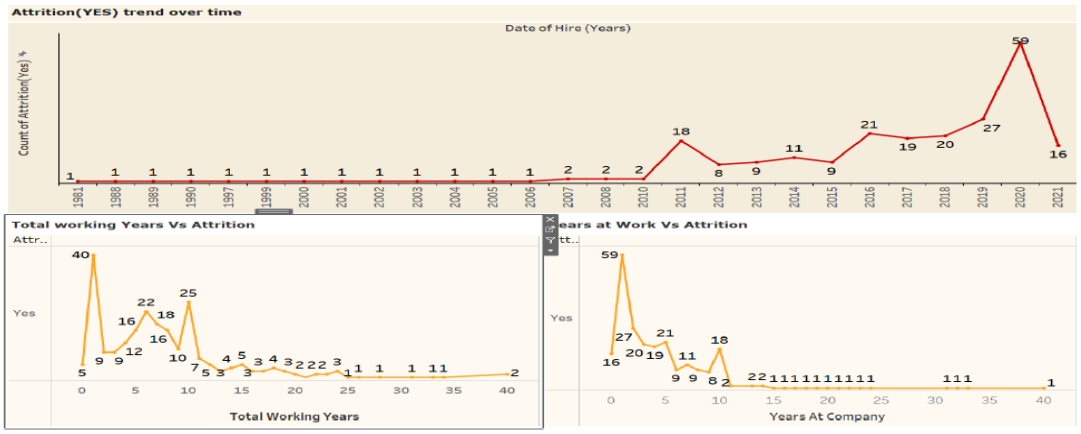
17 columns are numerical variable, and 13 columns are categorical variables, while 2 are empty columns, respectively.

12



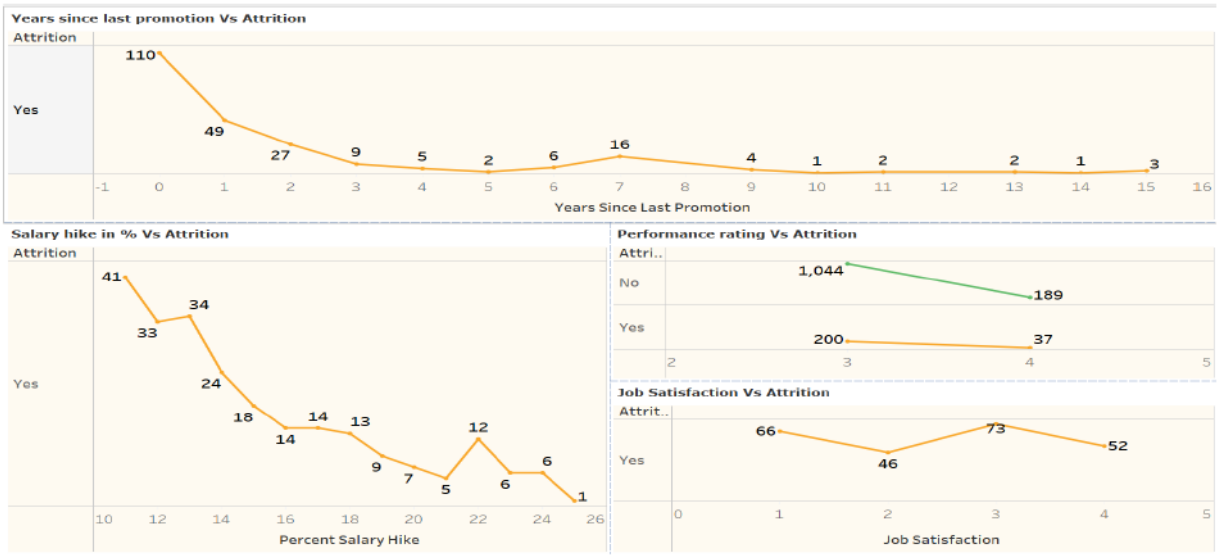
13

EXPLORATORY DATA ANALYSIS VISUALIZATIONS (TABLEAU) DASHBOARD 2



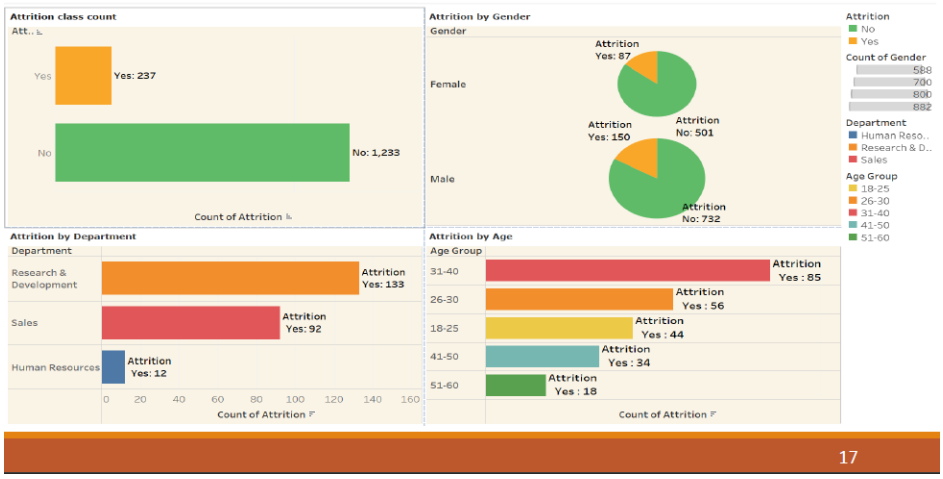
15

EXPLORATORY DATA ANALYSIS VISUALIZATIONS (TABLEAU) DASHBOARD 3



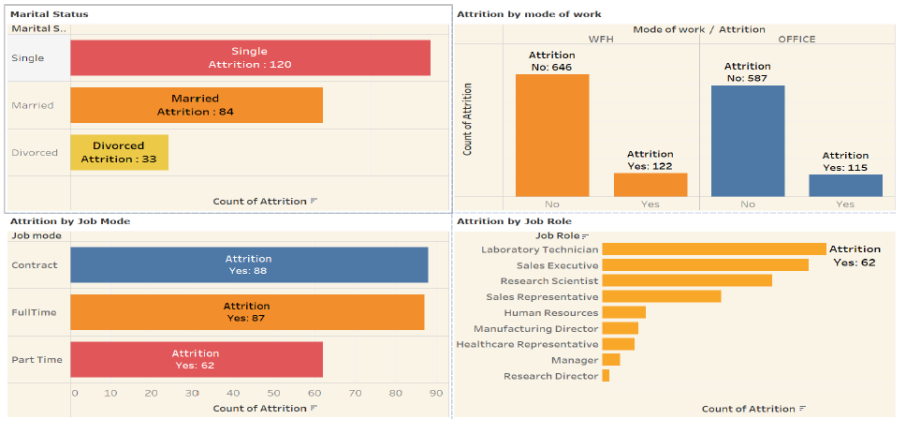
16

EXPLORATORY DATA ANALYSIS VISUALIZATIONS (TABLEAU) DASHBOARD 4



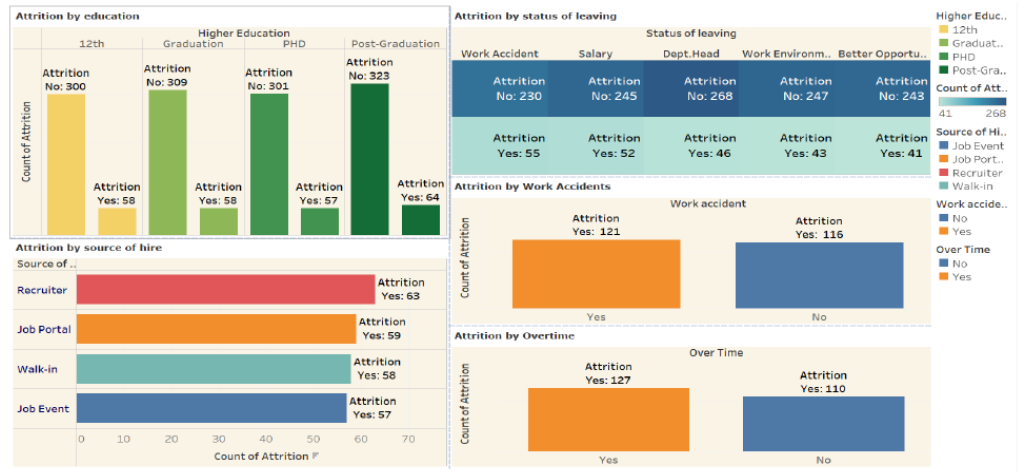
17

EXPLORATORY DATA ANALYSIS VISUALIZATIONS (TABLEAU) DASHBOARD 5



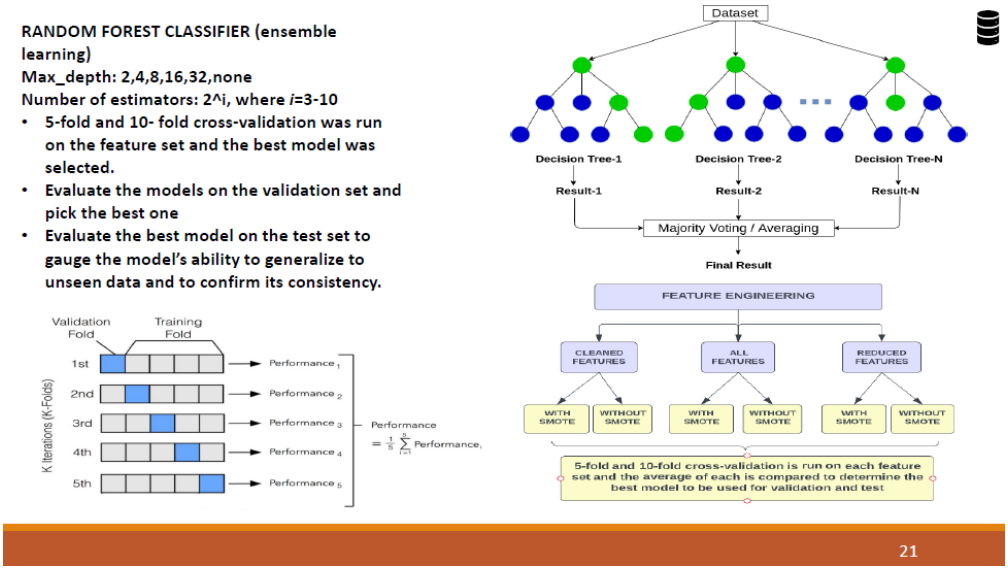
18

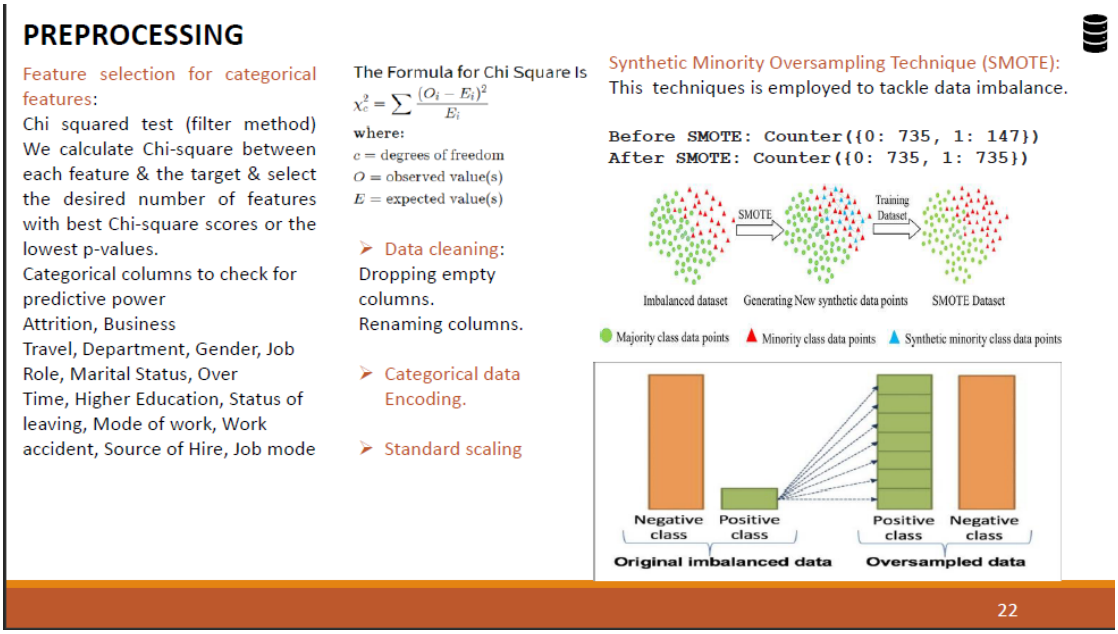
EXPLORATORY DATA ANALYSIS VISUALIZATIONS (TABLEAU) DASHBOARD 6



19

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
20


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.
Validation Fold
Training Fold
K Iterations (K-Folds)
1st
2nd
3rd
4th
5th
Performance
Dataset
Decision Tree-1
Decision Tree-2
Decision Tree-N
Result-1
Result-2
Result-N
Majority Voting / Averaging
Final Result
FEATURE ENGINEERING
CLEANED FEATURES
ALL FEATURES
REDUCED FEATURES
WITH SMOTE
WITHOUT SMOTE
5-fold and 10-fold cross-validation is run on each feature set and the average of each is compared to determine the best model to be used for validation and test
21


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})
SMOTE
Training Dataset
Imbalanced dataset
Generating New synthetic data points
SMOTE Dataset
Majority class data points
Minority class data points
Synthetic minority class data points
Negative class
Positive class
Positive class
Negative class
Original imbalanced data
Oversampled data
22

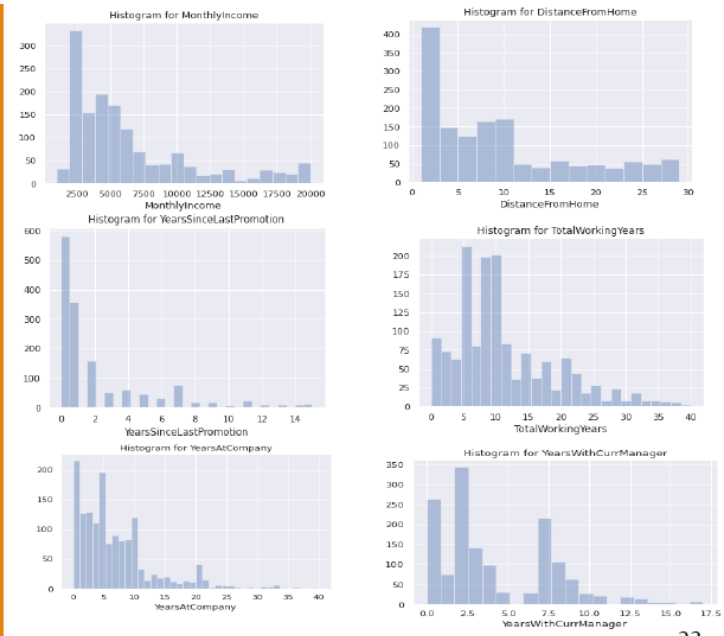
Data transformation:

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

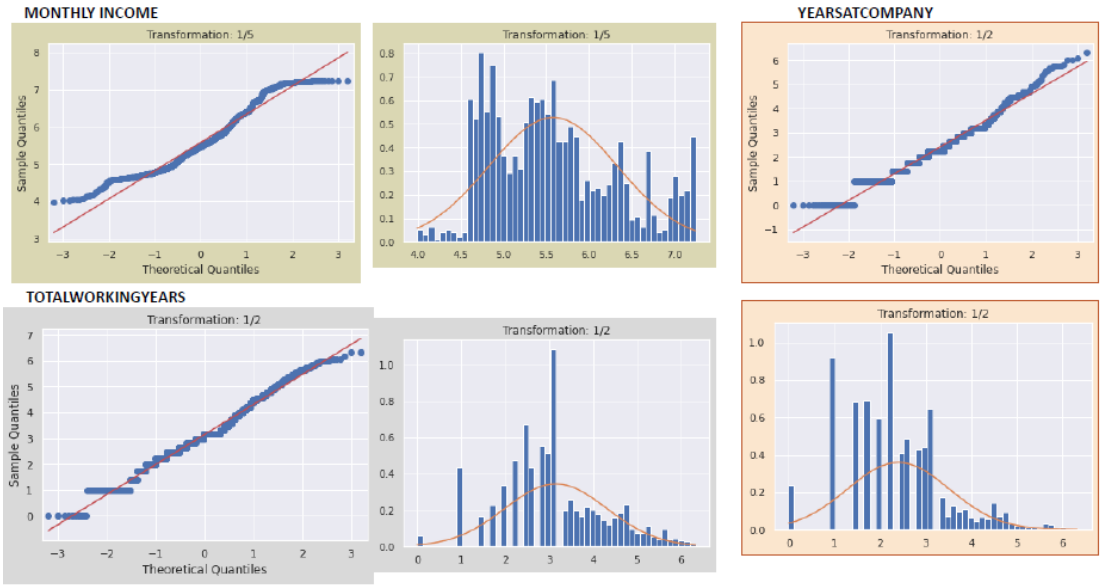
Columns to be transformed: Monthly Income, TotalWorkingYears, YearsAtCompany,

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

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



23



24

RESEARCH RESULTS

25

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

26

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

27

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

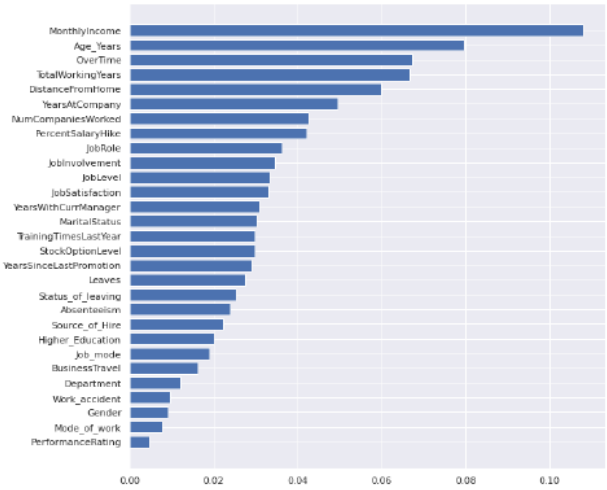
Feature set	ACC	PRECISION	RECALL	F1-SCORE	ROC_AUC	LATENCY
SMOTE CLEAN	0.878	0.652	0.349	0.455	0.658	49.5ms
SMOTE ALL	0.871	0.6	0.349	0.441	0.654	40.6ms

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

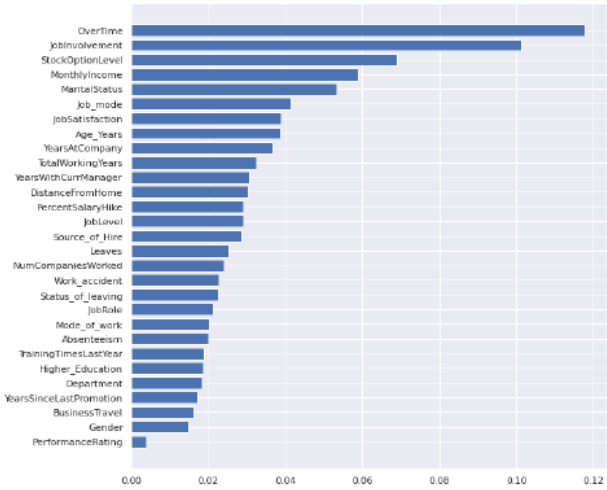
Feature set	ACC	PRECISION	RECALL	F1-SCORE	ROC_AUC	LATENCY
SMOTE CLEAN	0.864	0.565	0.302	0.394	0.631	36.7ms
SMOTE ALL	0.867	0.577	0.349	0.435	0.653	16.3ms

28

CLEANED FEATURE IMPORTANCE PLOT

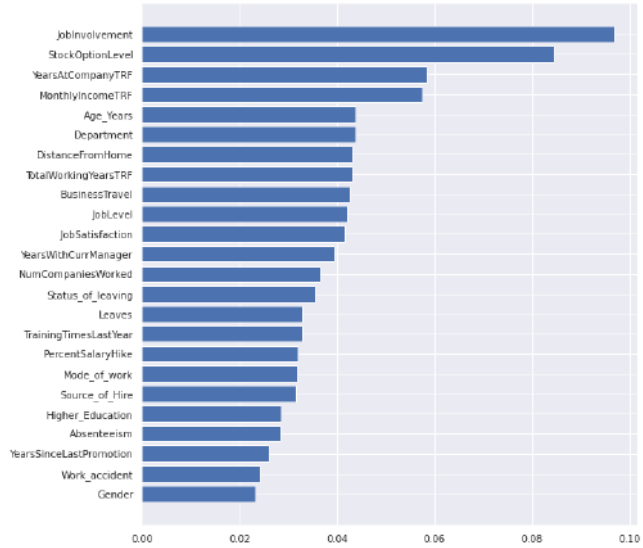


SMOTE CLEANED FEATURE PLOT

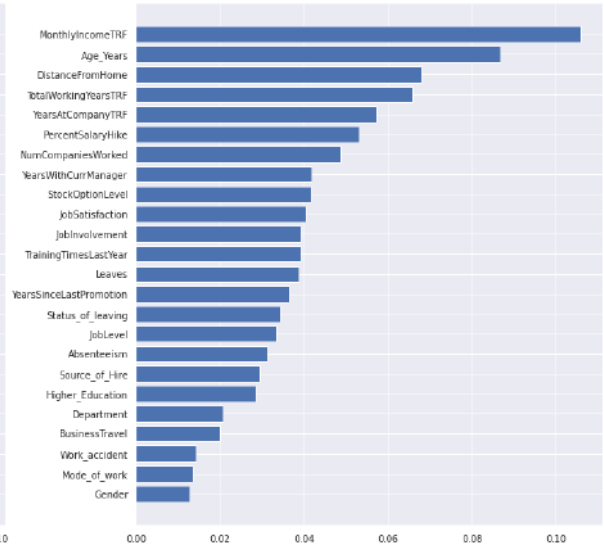


29

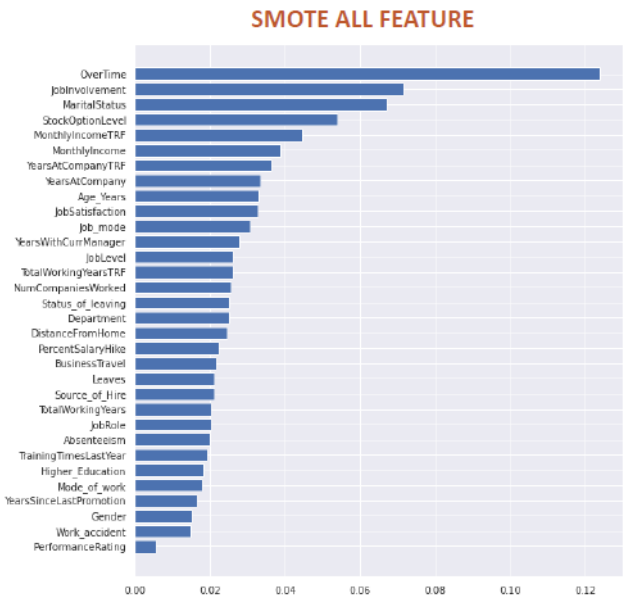
SMOTE REDUCED FEATURE

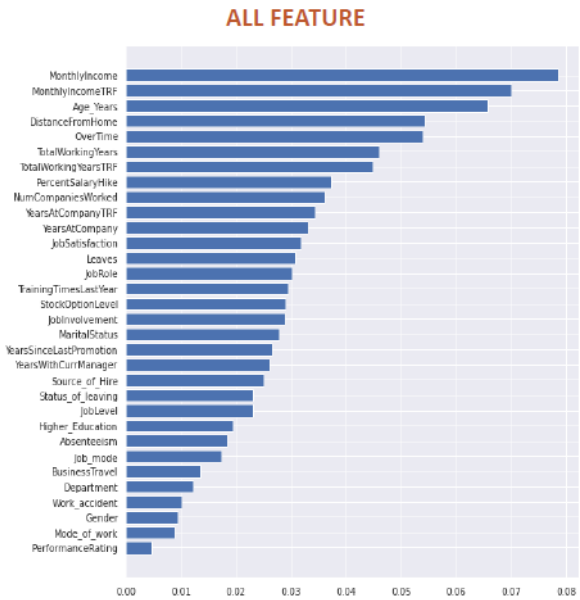


REDUCED FEATURE



30





31

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.

32

CHALLENGES, IN PROGRESS & FUTURE WORK

33

- 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.

34

THANK YOU

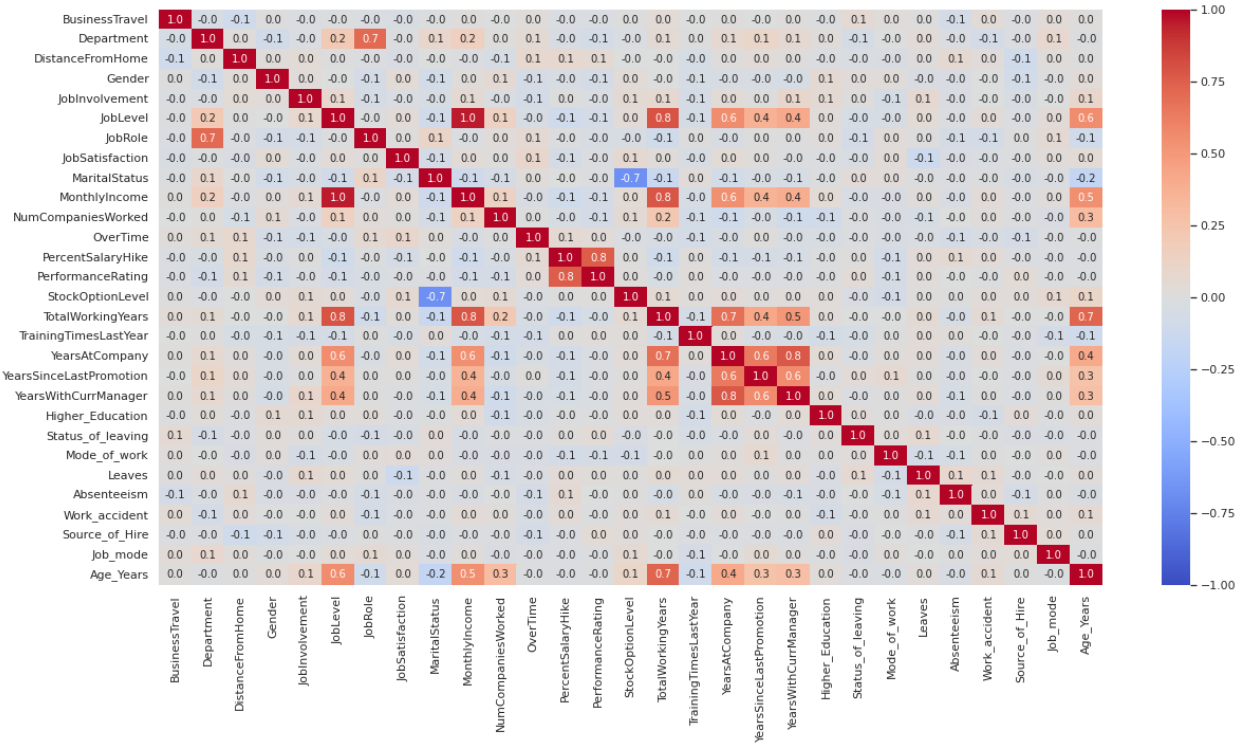
36

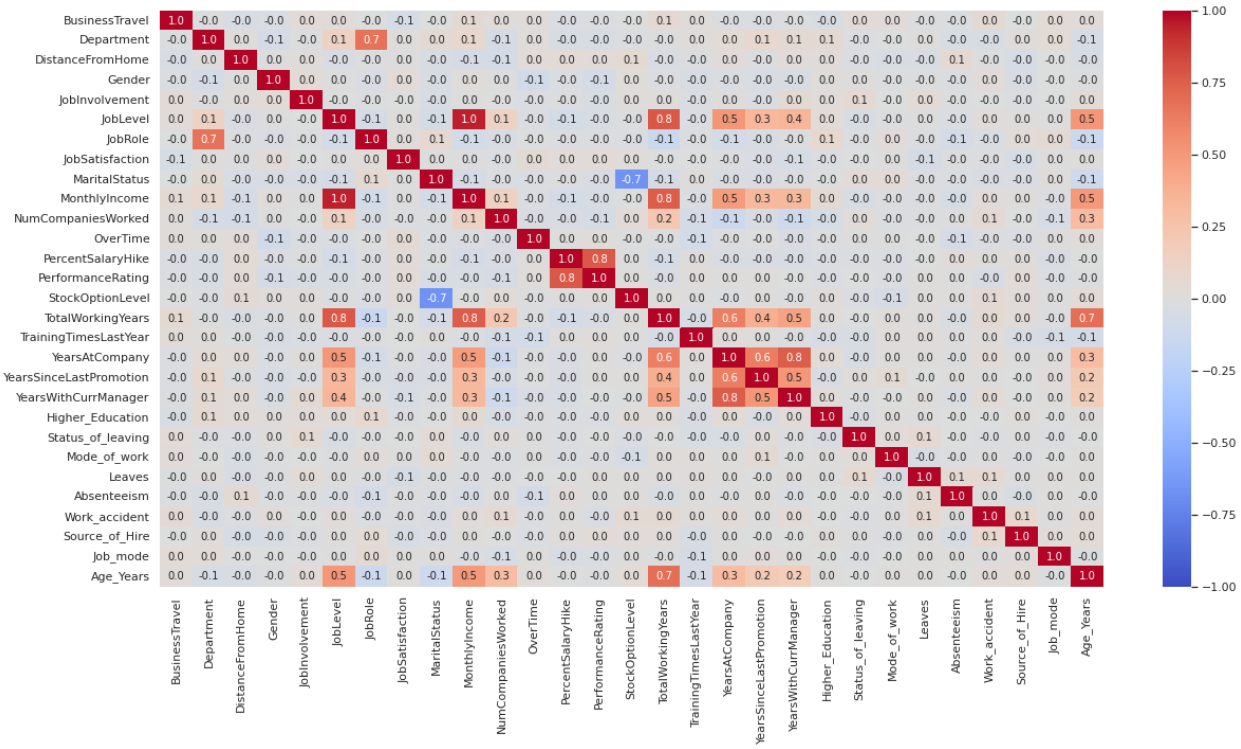
Appendix C –Meeting Records

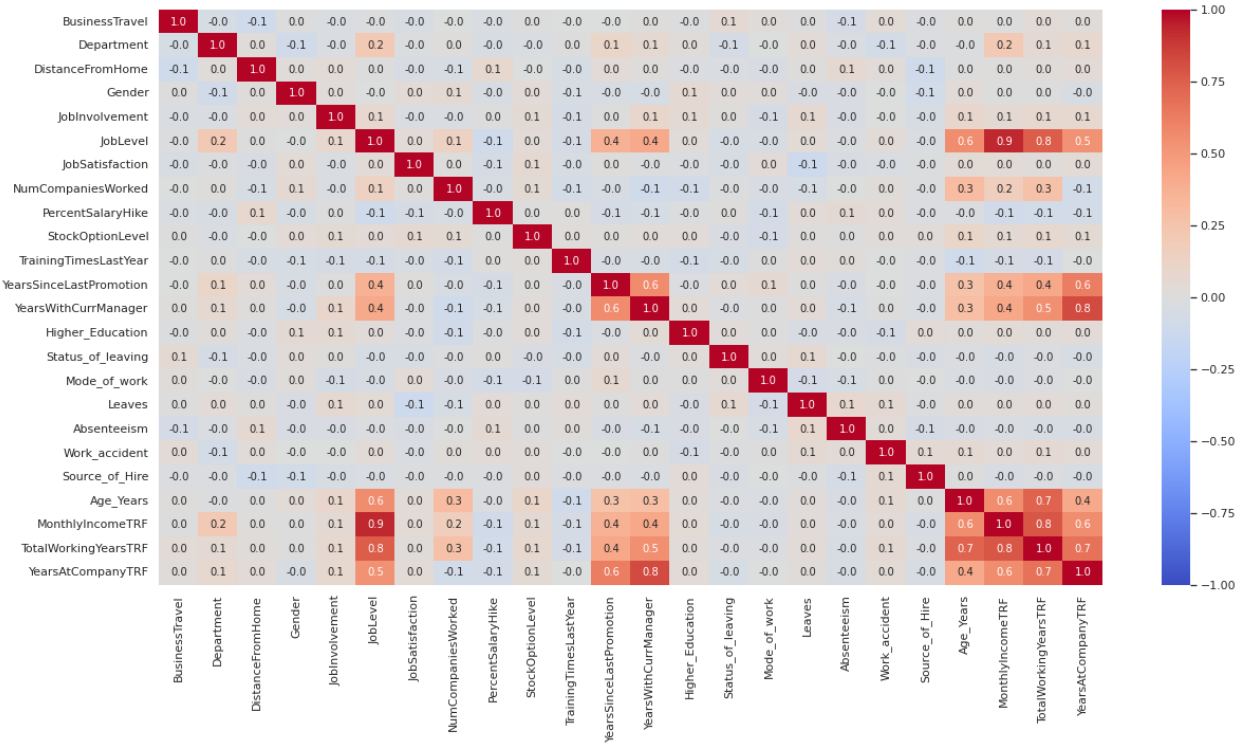
Meeting	Discussion	Date	Time
1.	• Introduction • Talked about compulsory meeting checkpoints every fortnight, time, and location • Suggested project topics	17/05/2022	06:00pm-06:30pm
2.	• Shared materials and videos on machine learning via e-mail to help prepare my mind for the project. Completed linkedin course recommended by supervisor	18/05/2022	
3.	• Shared project draft and plan with supervisor • Discussed on how to kick start on machine learning task, feature engineering and implementation.	30/05/2022	
4.	• Discussion on choice of dataset and preparation of project proposal • Ibm attrition data set chosen and approved by supervisor	01/06/2022	
5.	• Discussion on choice of programming language and approach to solving problem. • Ethical application completion and submission.	10/06/2022	
6.	• Discussed methodology, problem statement, aims and objectives of project.	10/06/2022	06:00pm-06:30pm
7.	• Ethical application approved • Discussed further techniques like feature selection, feature engineering, data visualization tools and shared progress on project.	08/07/2022	06:00pm-06:30pm
8.	• Share source code via email with supervisor. • Discussion on further technique to ensure quality of research project • Shared challenges concerning coding with python on google colab. • Choice of algorithms to use.	21/07/2022	06:00pm-06:30
9.	• Discussion about presentation and 2nd		

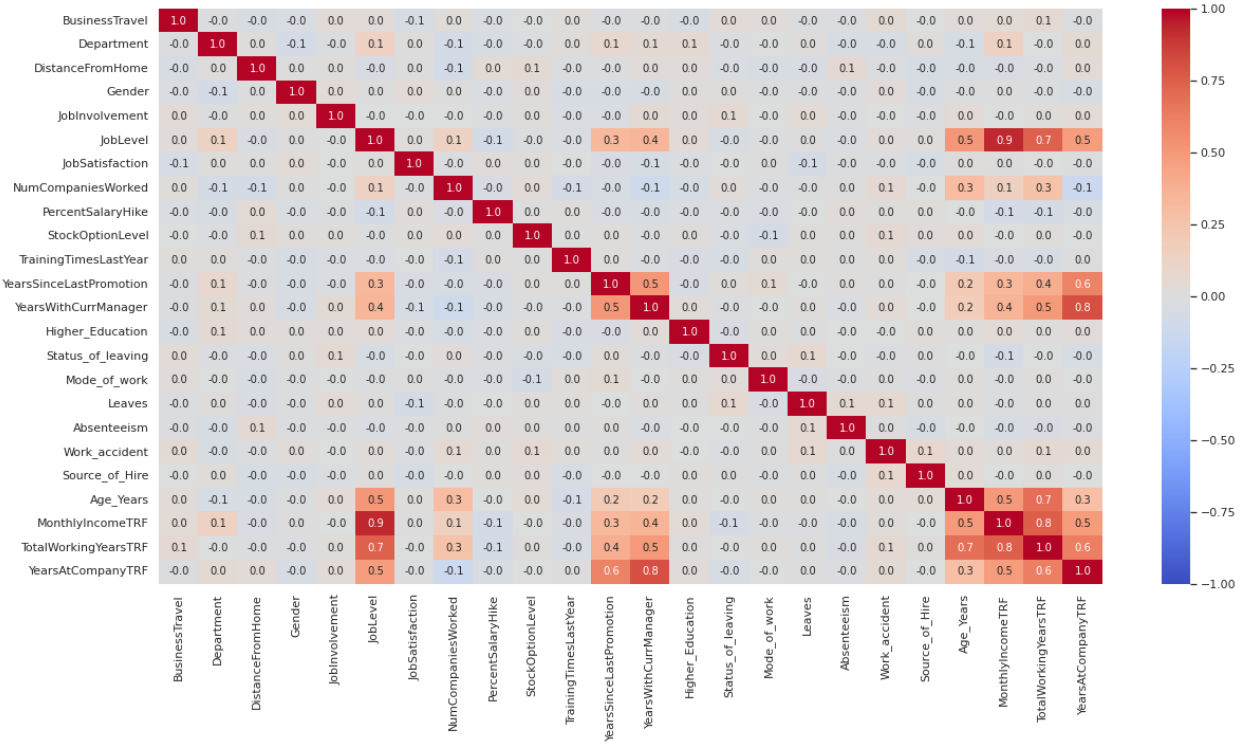
	supervisor allocated. • Progress discussion • Adviced me to state reason in report why I choose random forest classifiers over other algorithms.	26/07/2022	06:00-06:30pm
10.	• Gave tips on marking rubic • Shared dates for presentation • Booked appointment for project presentation	03/08/2022	
11.	• Project presentation • Discussed on what to do when reporting project. Reiterated what I should include in project based on the cw3 outline.	11/08/2022	

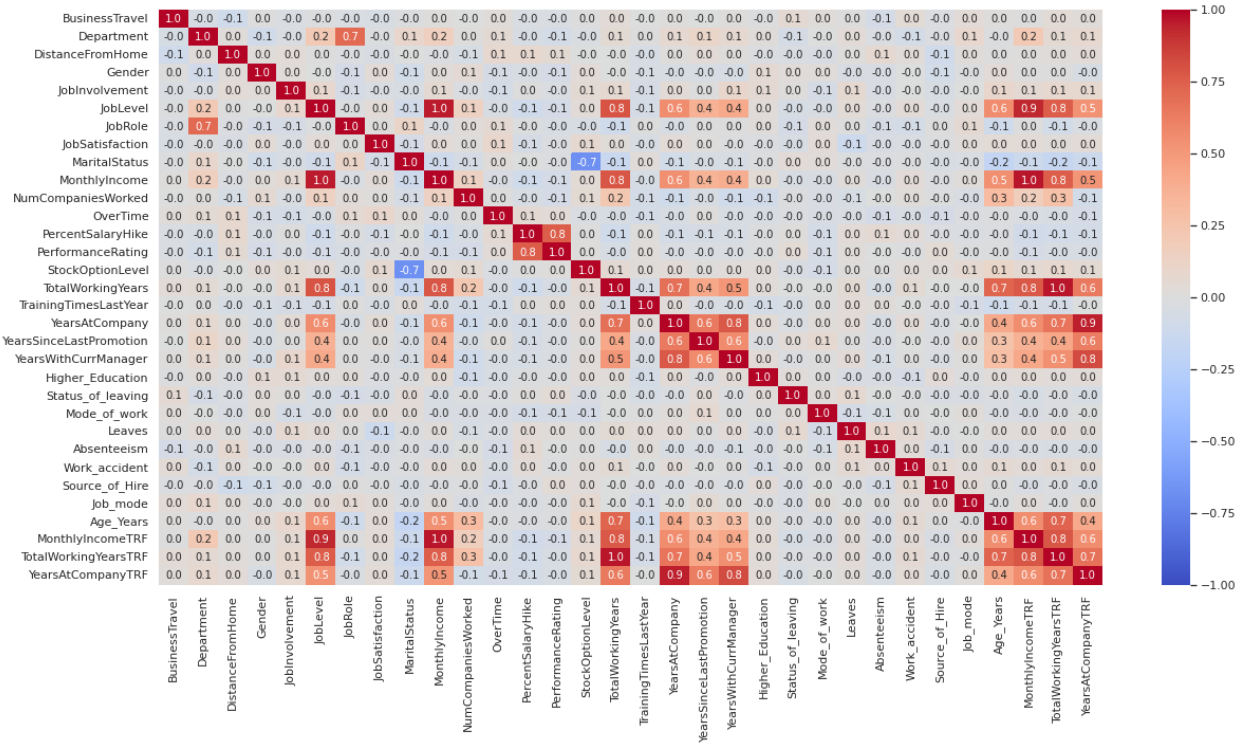
Appendix X – visualizations

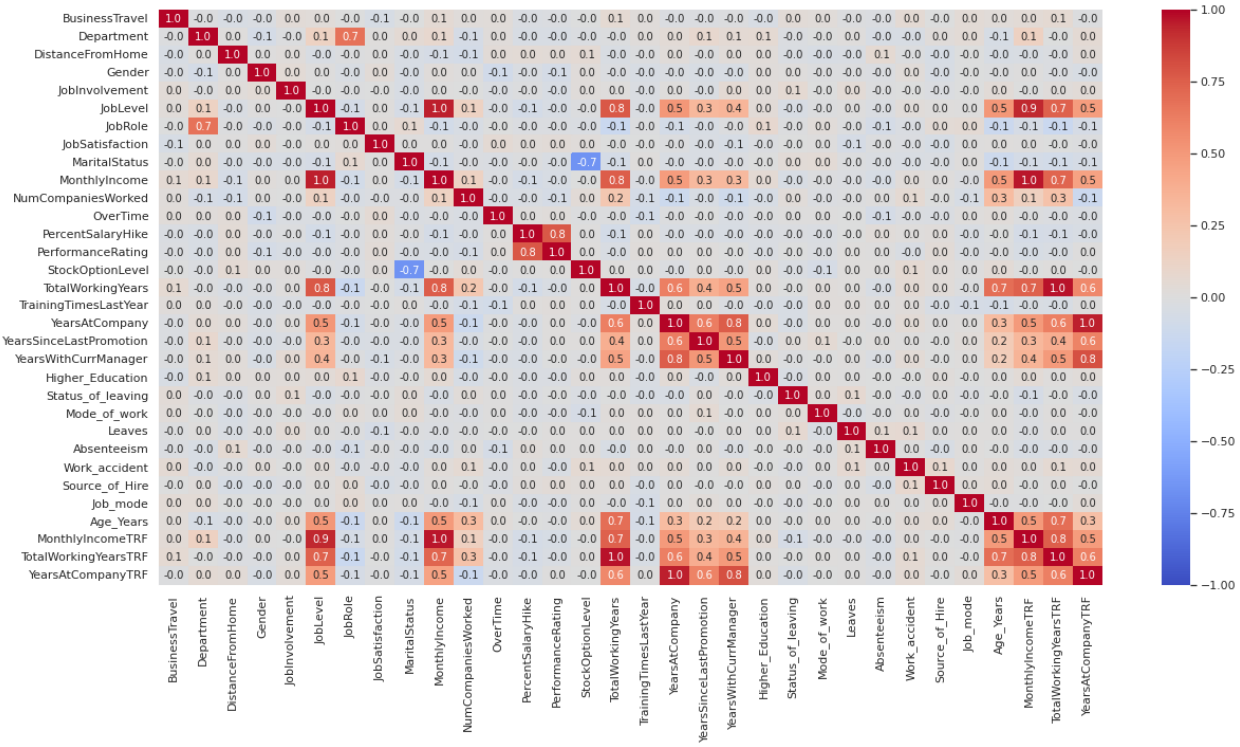


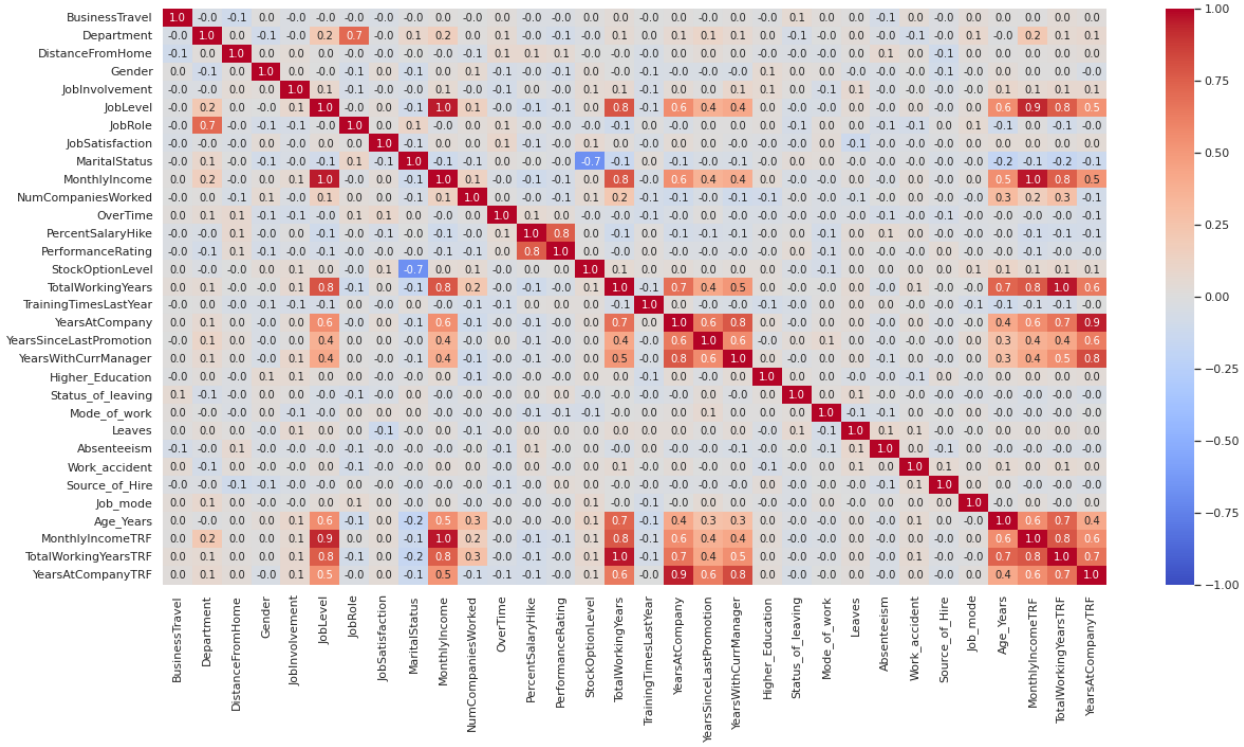


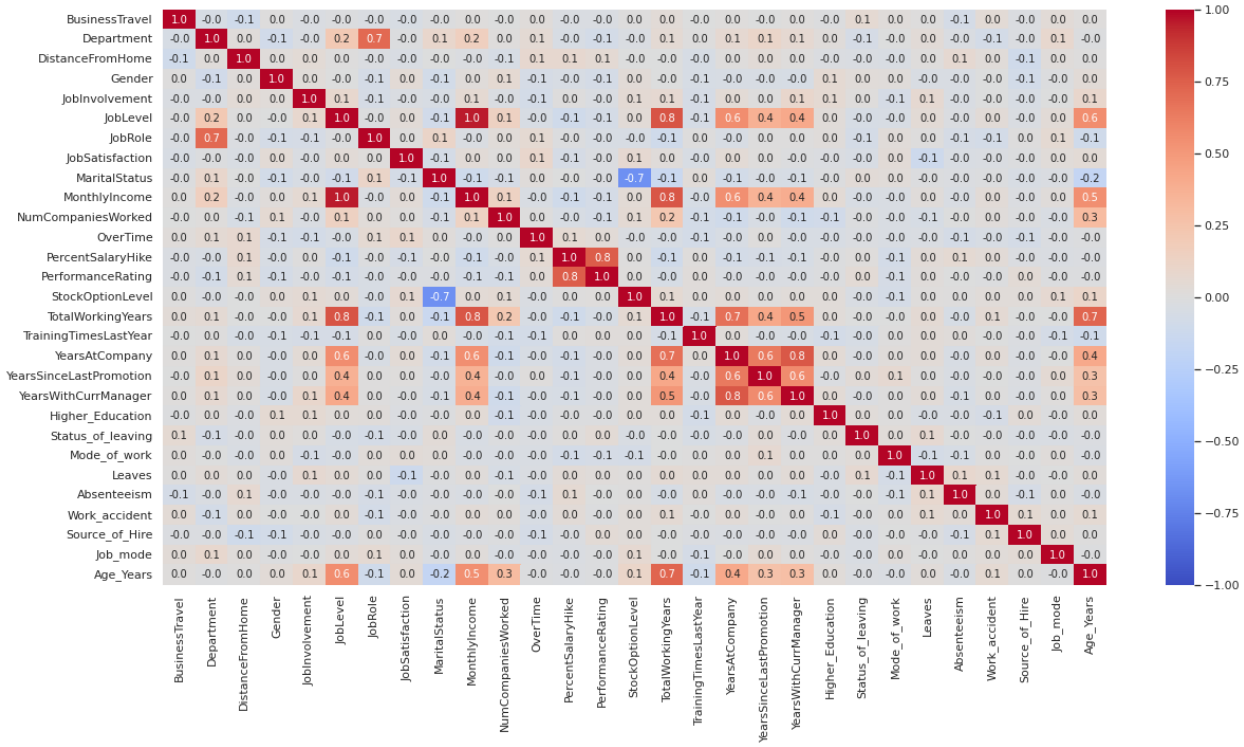

BusinessTravel
Department
DistanceFromHome
Gender
JobInvolvement
JobLevel
JobSatisfaction
NumCompaniesWorked
PercentSalaryHike
StockOptionLevel
TrainingTimesLastYear
YearsSinceLastPromotion
YearsWithCurrManager
Higher_Education
Status_of_leaving
Mode_of_work
Leaves
Absenteeism
Work_accident
Source_of_Hire
Age_Years
MonthlyIncomeTRF
TotalWorkingYearsTRF
YearsAtCompanyTRF
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

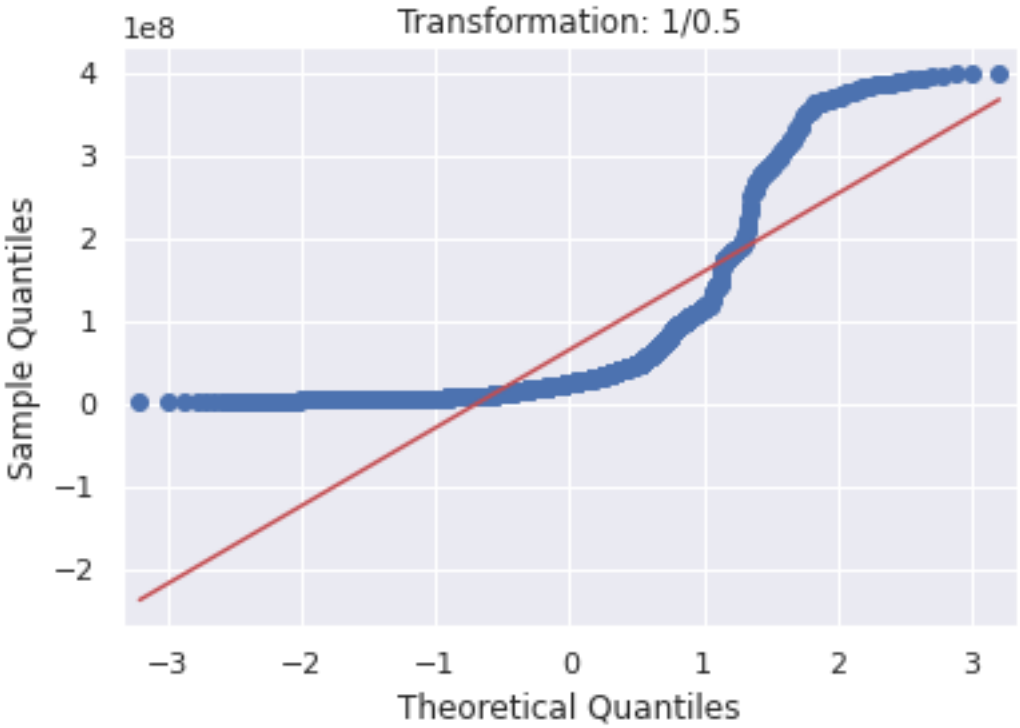

BusinessTravel
Department
DistanceFromHome
Gender
JobInvolvement
JobLevel
JobSatisfaction
NumCompaniesWorked
PercentSalaryHike
StockOptionLevel
TrainingTimesLastYear
YearsSinceLastPromotion
YearsWithCurrManager
Higher_Education
Status_of_leaving
Mode_of_work
Leaves
Absenteeism
Work_accident
Source_of_Hire
Age_Years
MonthlyIncomeTRF
TotalWorkingYearsTRF
YearsAtCompanyTRF
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


BusinessTravel
Department
DistanceFromHome
Gender
JobInvolvement
JobLevel
JobRole
JobSatisfaction
MaritalStatus
MonthlyIncome
NumCompaniesWorked
OverTime
PercentSalaryHike
PerformanceRating
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
YearsAtCompany
YearsSinceLastPromotion
YearsWithCurrManager
Higher_Education
Status_of_leaving
Mode_of_work
Leaves
Absenteeism
Work_accident
Source_of_Hire
Job_mode
Age_Years
MonthlyIncomeTRF
TotalWorkingYearsTRF
YearsAtCompanyTRF
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

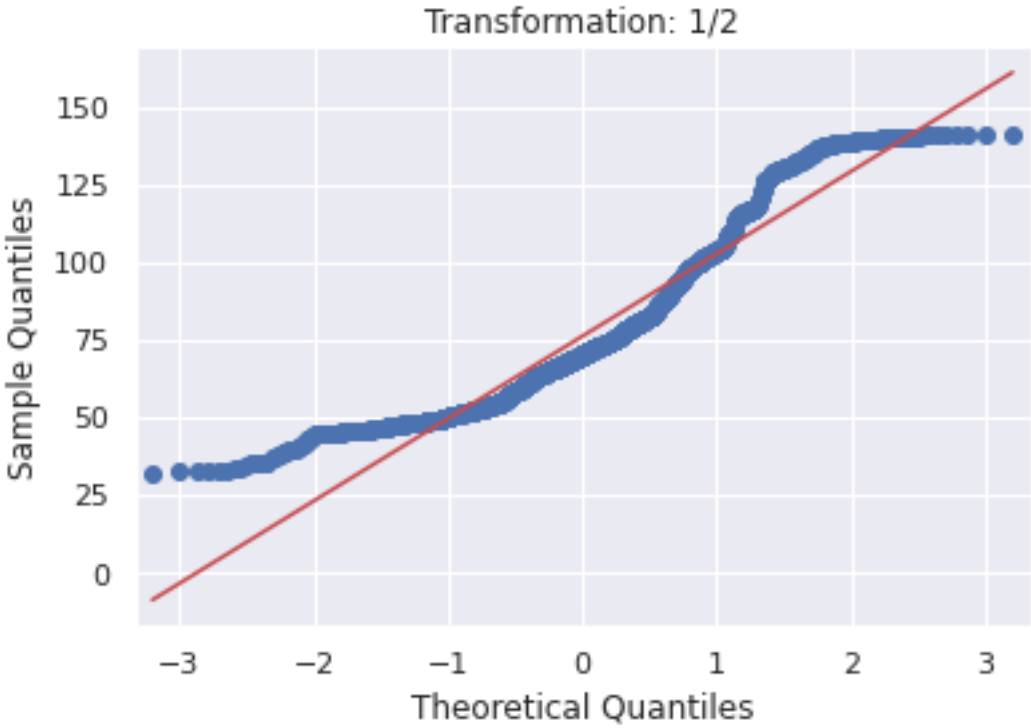

BusinessTravel
Department
DistanceFromHome
Gender
JobInvolvement
JobLevel
JobRole
JobSatisfaction
MaritalStatus
MonthlyIncome
NumCompaniesWorked
OverTime
PercentSalaryHike
PerformanceRating
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
YearsAtCompany
YearsSinceLastPromotion
YearsWithCurrManager
Higher_Education
Status_of_leaving
Mode_of_work
Leaves
Absenteeism
Work_accident
Source_of_Hire
Job_mode
Age_Years
MonthlyIncomeTRF
TotalWorkingYearsTRF
YearsAtCompanyTRF
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

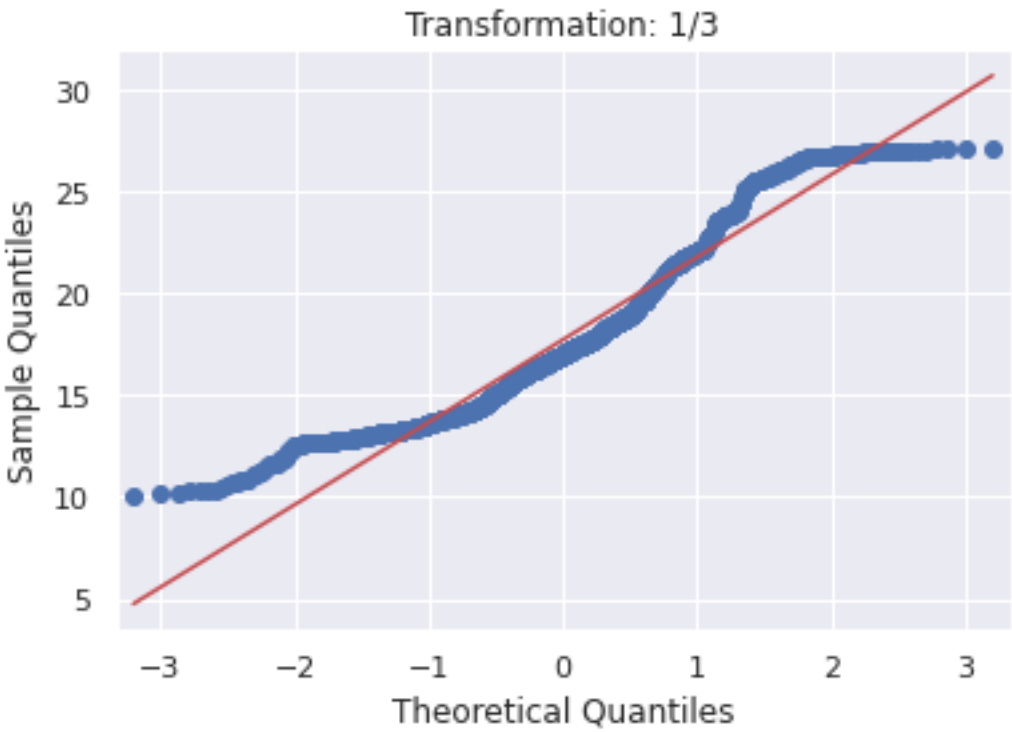

BusinessTravel
Department
DistanceFromHome
Gender
JobInvolvement
JobLevel
JobRole
JobSatisfaction
MaritalStatus
MonthlyIncome
NumCompaniesWorked
OverTime
PercentSalaryHike
PerformanceRating
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
YearsAtCompany
YearsSinceLastPromotion
YearsWithCurrManager
Higher_Education
Status_of_leaving
Mode_of_work
Leaves
Absenteeism
Work_accident
Source_of_Hire
Job_mode
Age_Years
MonthlyIncomeTRF
TotalWorkingYearsTRF
YearsAtCompanyTRF
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

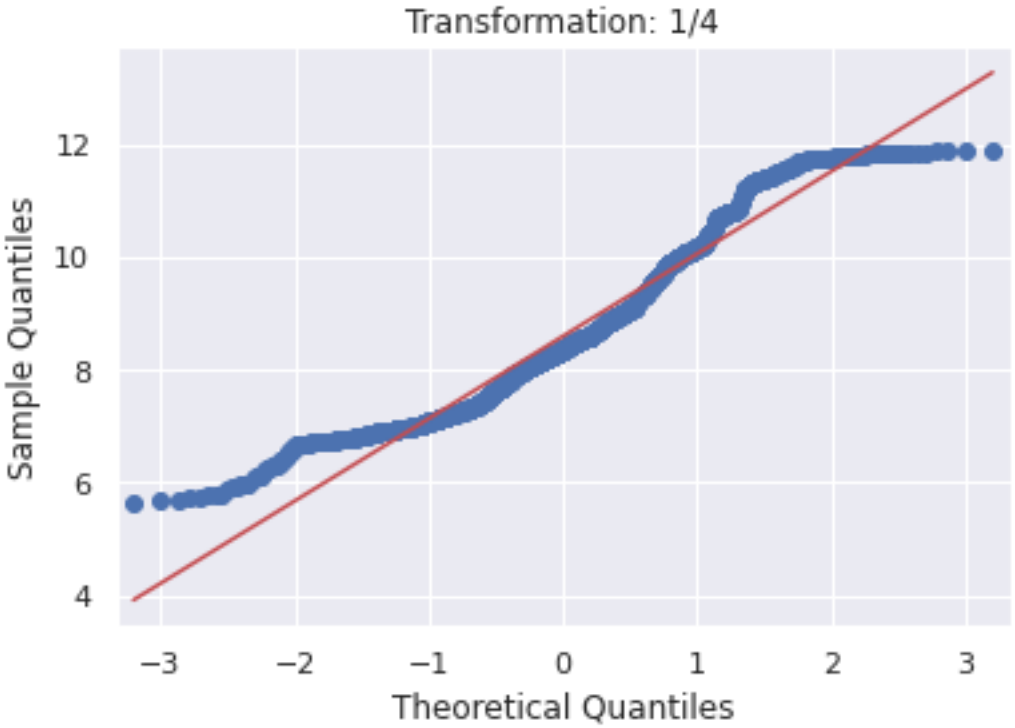

BusinessTravel
Department
DistanceFromHome
Gender
JobInvolvement
JobLevel
JobRole
JobSatisfaction
MaritalStatus
MonthlyIncome
NumCompaniesWorked
OverTime
PercentSalaryHike
PerformanceRating
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
YearsAtCompany
YearsSinceLastPromotion
YearsWithCurrManager
Higher_Education
Status_of_leaving
Mode_of_work
Leaves
Absenteeism
Work_accident
Source_of_Hire
Job_mode
Age_Years
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

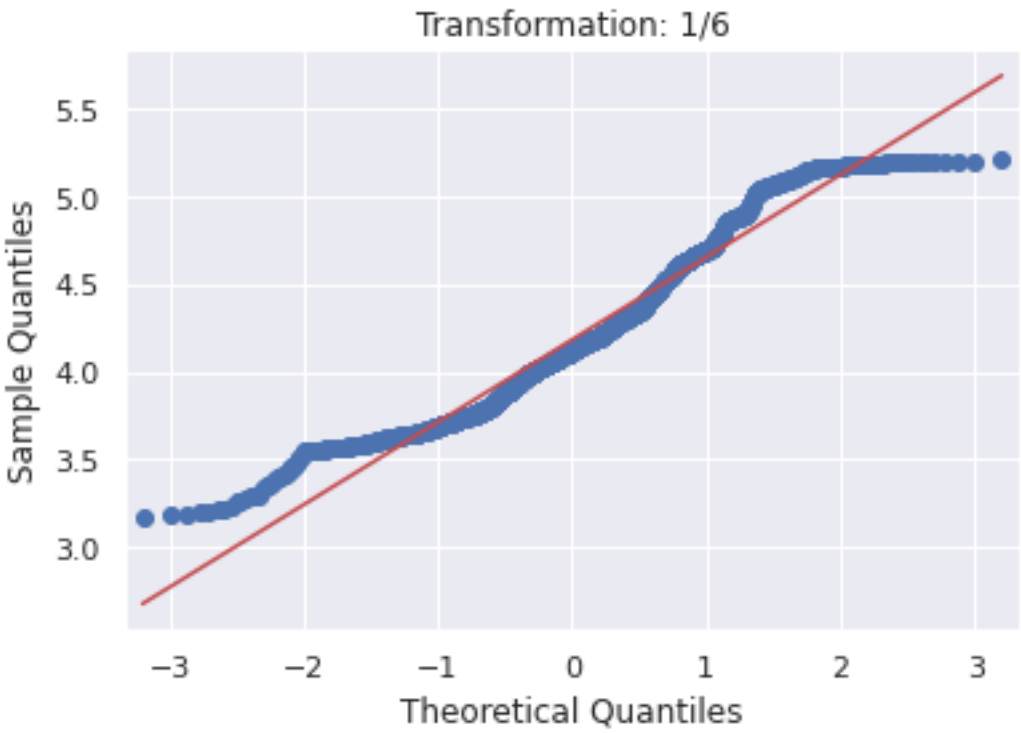

Transformation: 1/0.5
1e8
4
3
2
1
0
−1
−2
Sample Quantiles
−3
−2
−1
0
1
2
3
Theoretical Quantiles

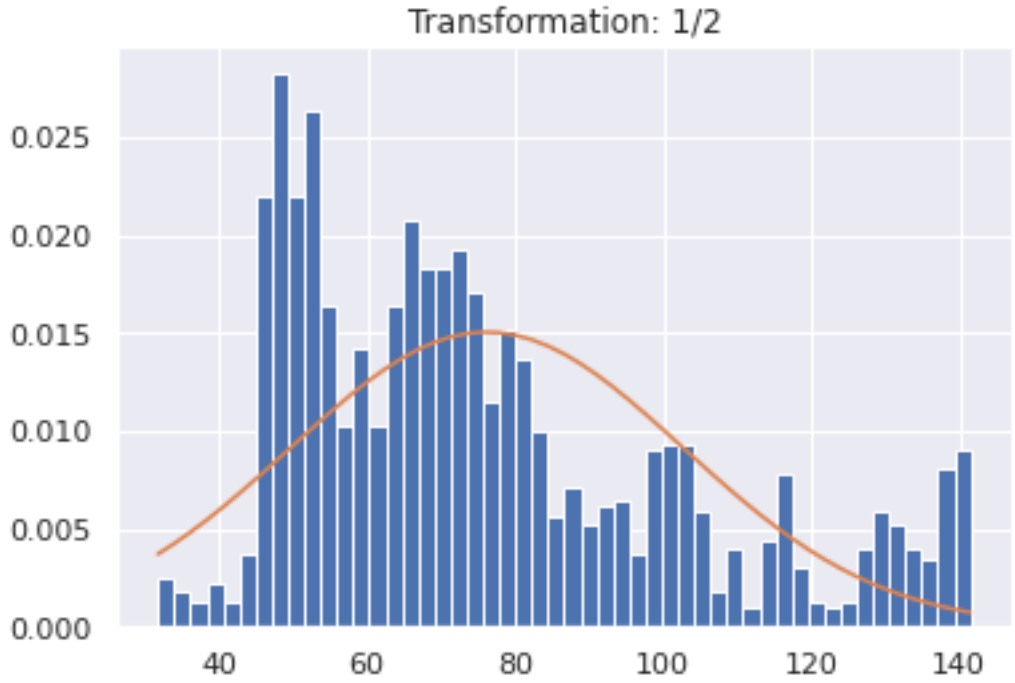
Transformation: 1/1
20000
15000
10000
5000
0
−5000
−10000
Sample Quantiles
−3
−2
−1
0
1
2
3
Theoretical Quantiles

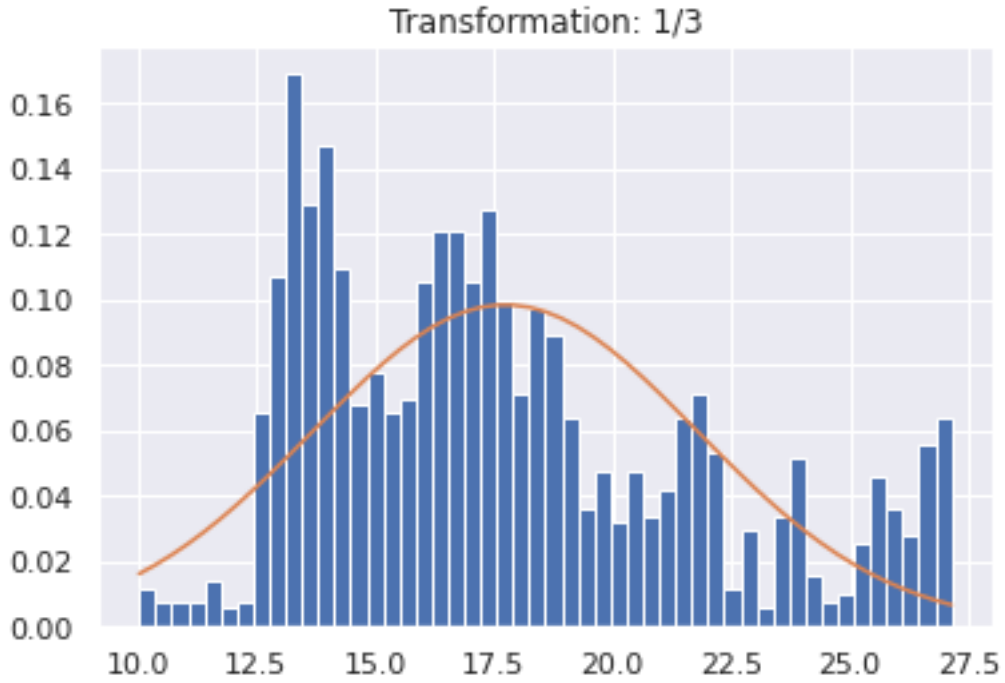

Transformation: 1/2
Sample Quantiles
150
125
100
75
50
25
0
−3
−2
−1
0
1
2
3
Theoretical Quantiles

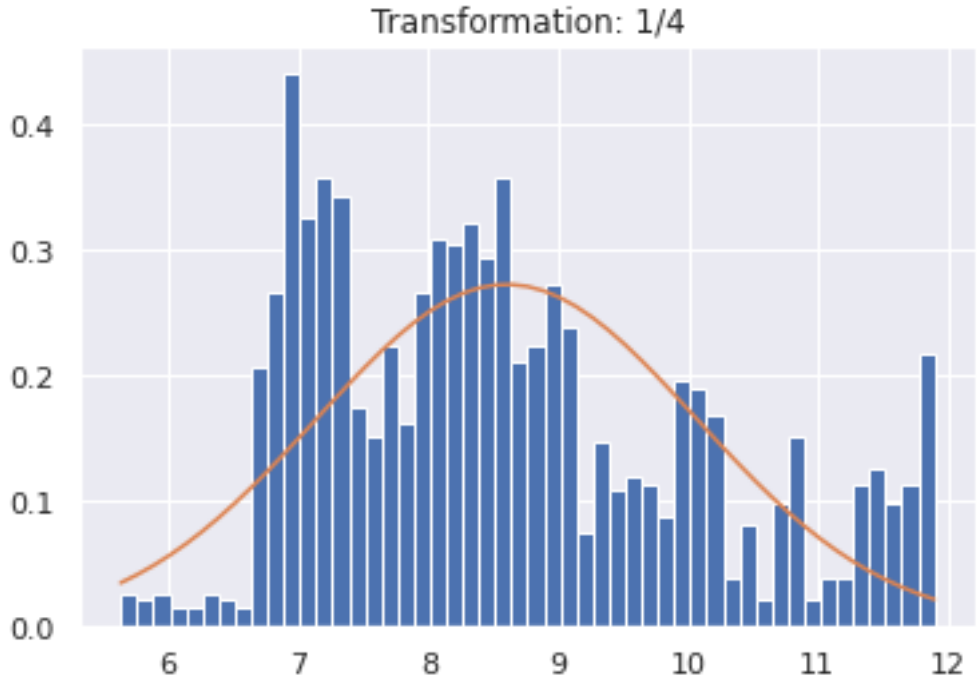

Transformation: 1/3
Sample Quantiles
30
25
20
15
10
5
−3
−2
−1
0
1
2
3
Theoretical Quantiles

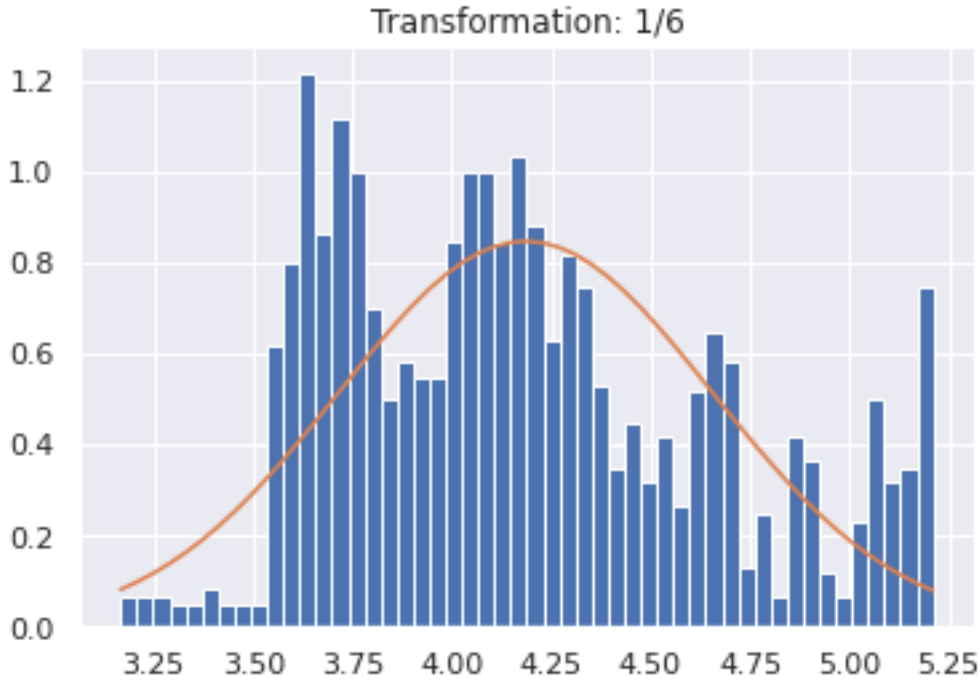

Transformation: 1/4
Sample Quantiles
12
10
8
6
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles

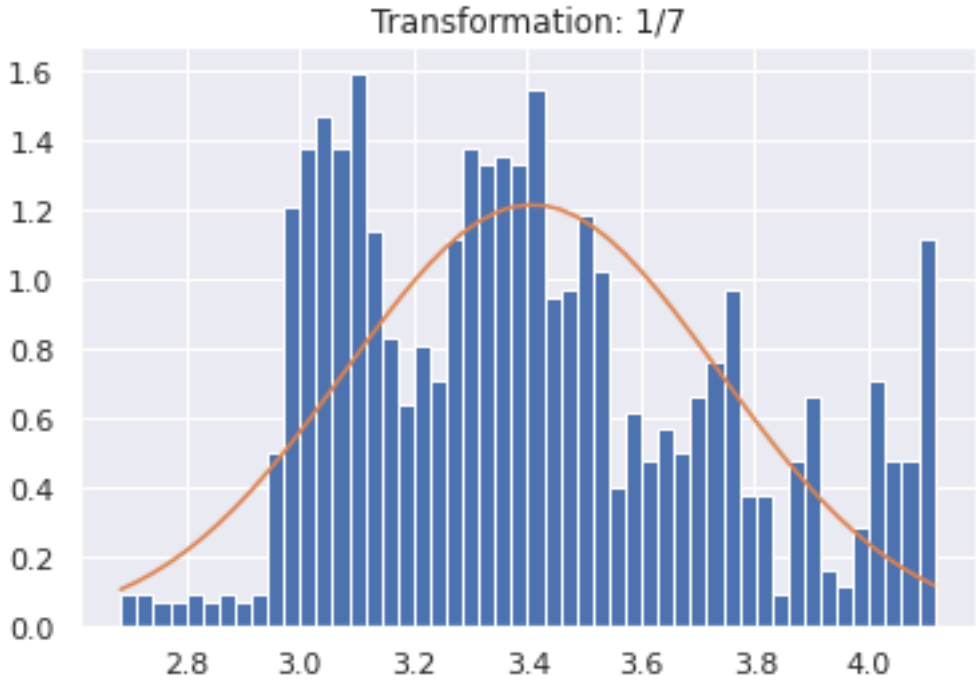

Transformation: 1/6
Sample Quantiles
5.5
5.0
4.5
4.0
3.5
3.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles

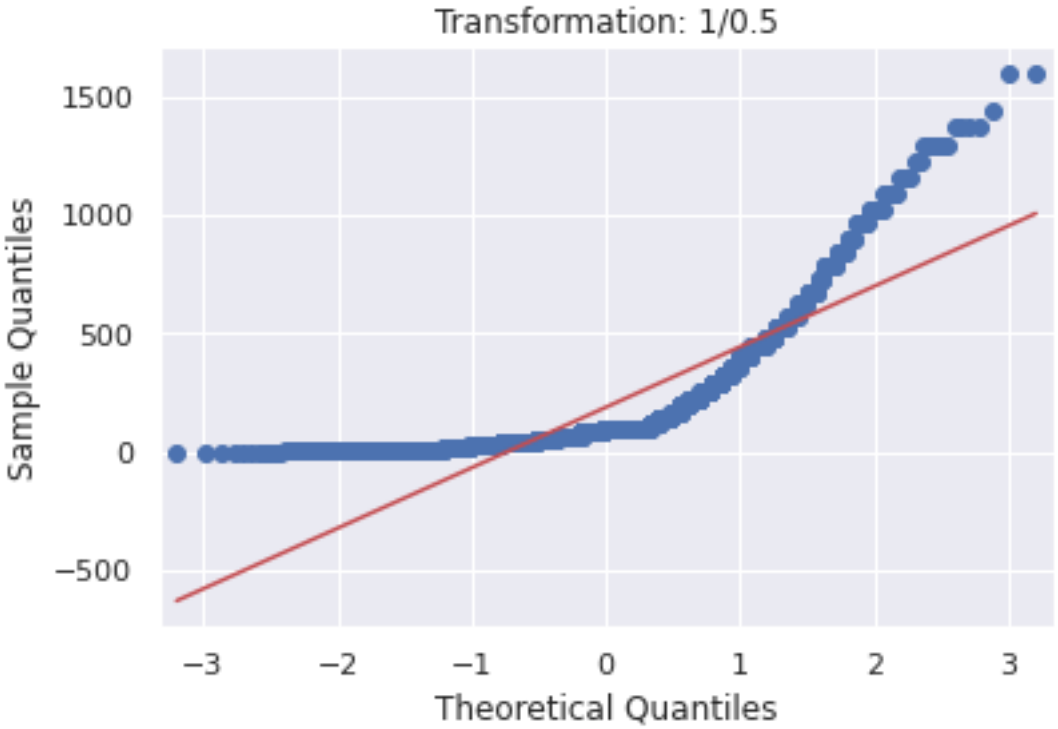

Transformation: 1/2
0.025
0.020
0.015
0.010
0.005
0.000
40
60
80
100
120
140

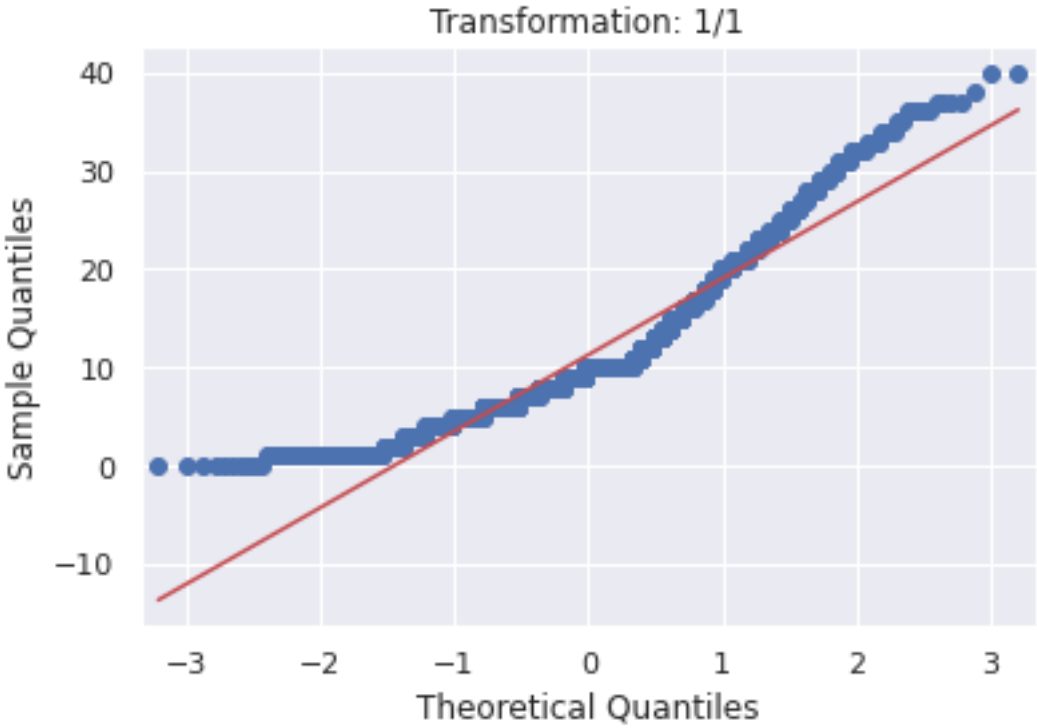

Transformation: 1/3
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
10.0
12.5
15.0
17.5
20.0
22.5
25.0
27.5

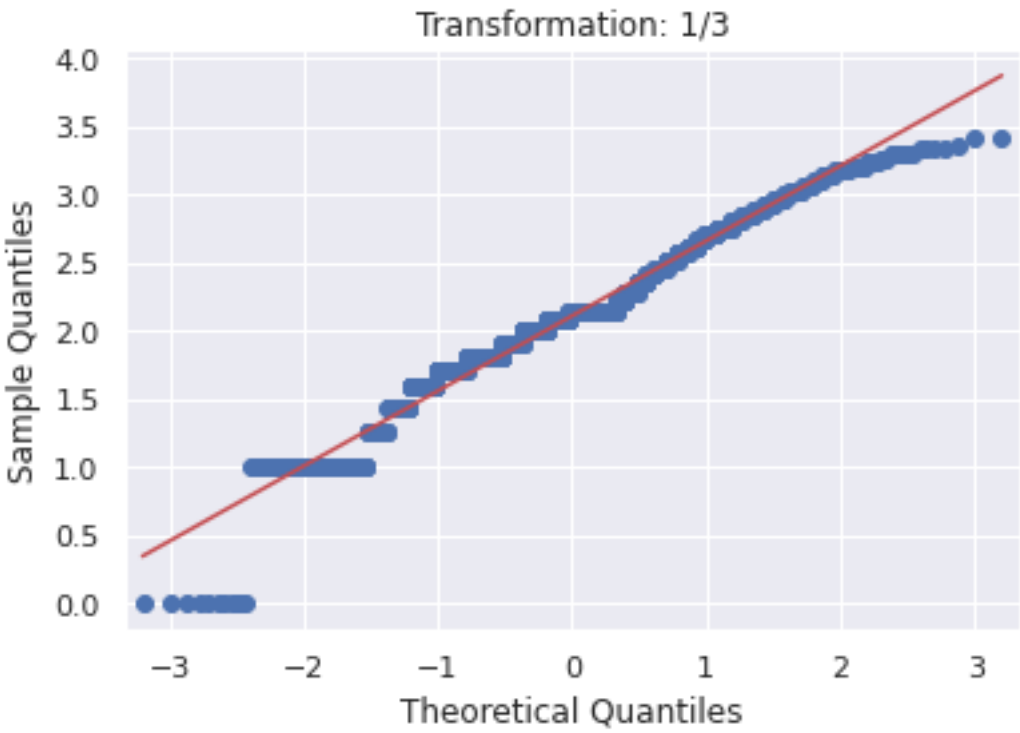

Transformation: 1/4
0.4
0.3
0.2
0.1
0.0
6
7
8
9
10
11
12

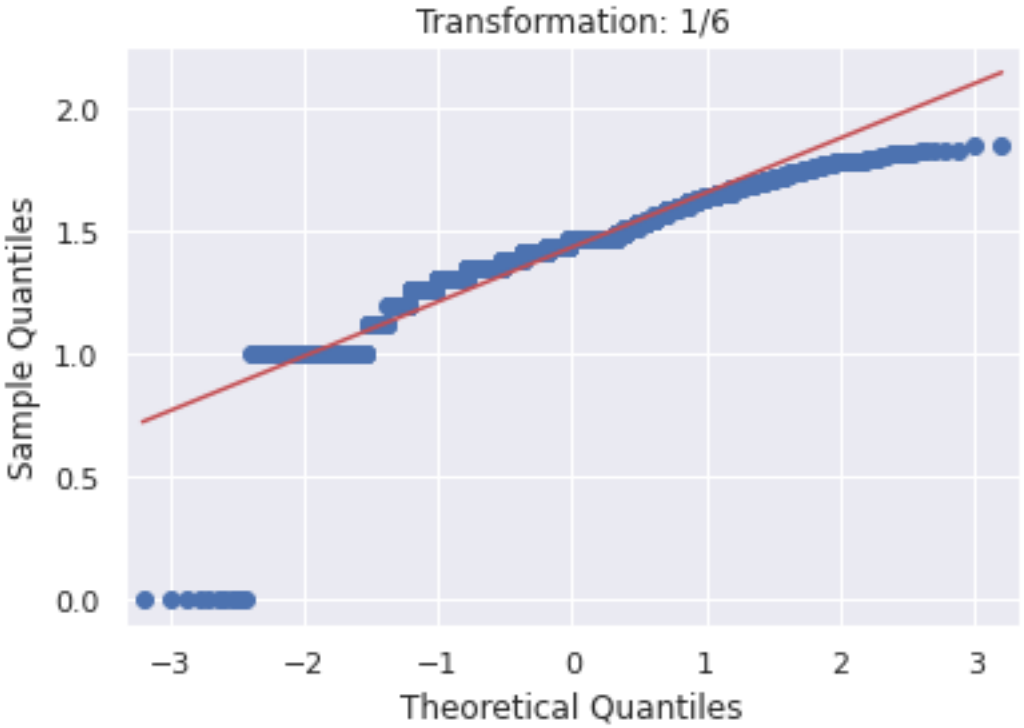

Transformation: 1/6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
3.25
3.50
3.75
4.00
4.25
4.50
4.75
5.00
5.25

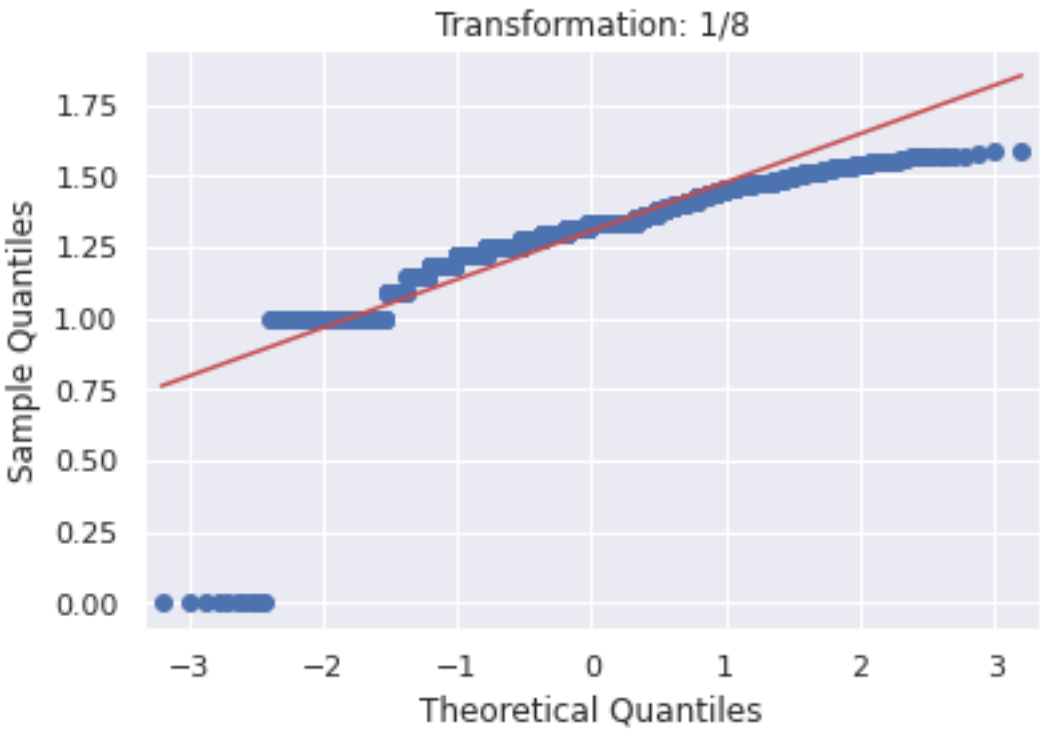

Transformation: 1/7
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
2.8
3.0
3.2
3.4
3.6
3.8
4.0

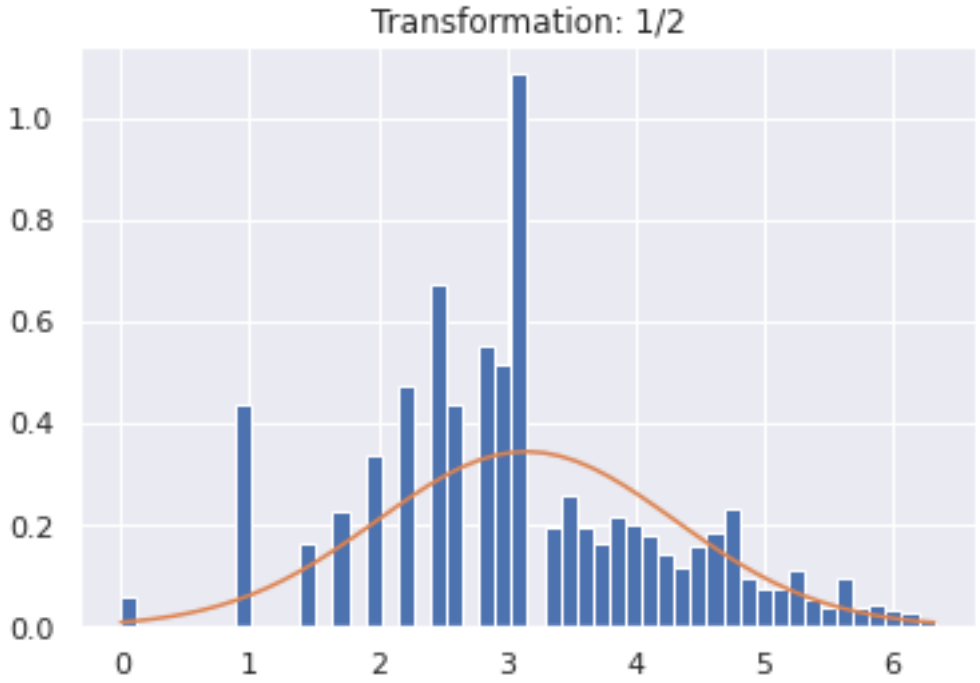

Transformation: 1/0.5
Sample Quantiles
1500
1000
500
0
−500
−3
−2
−1
0
1
2
3
Theoretical Quantiles

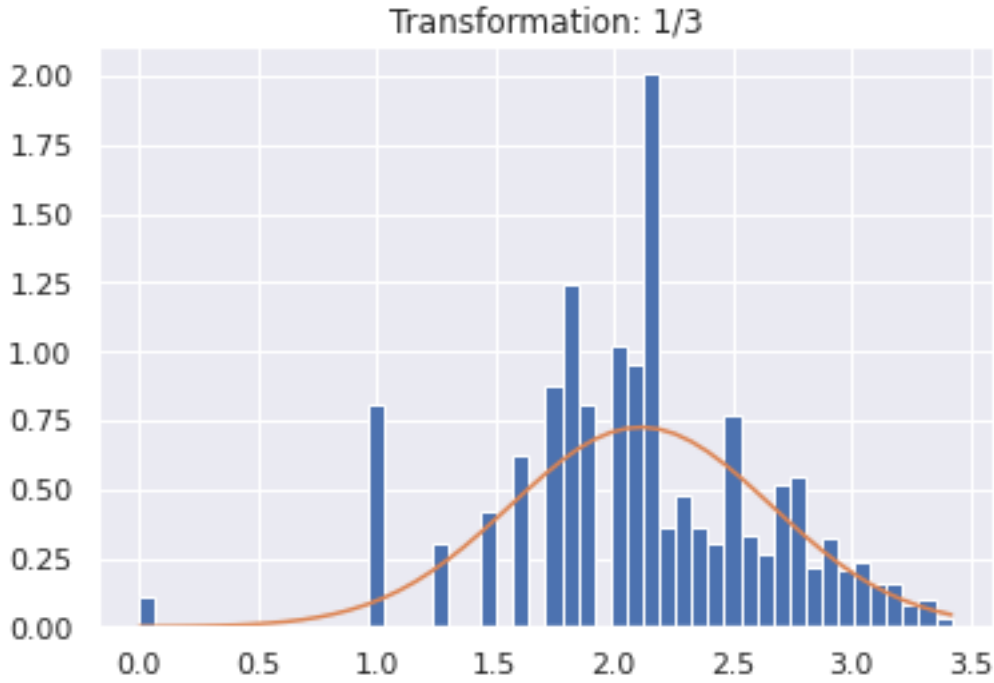

Transformation: 1/1
Sample Quantiles
Theoretical Quantiles
40
30
20
10
0
−10
−3
−2
−1
0
1
2
3

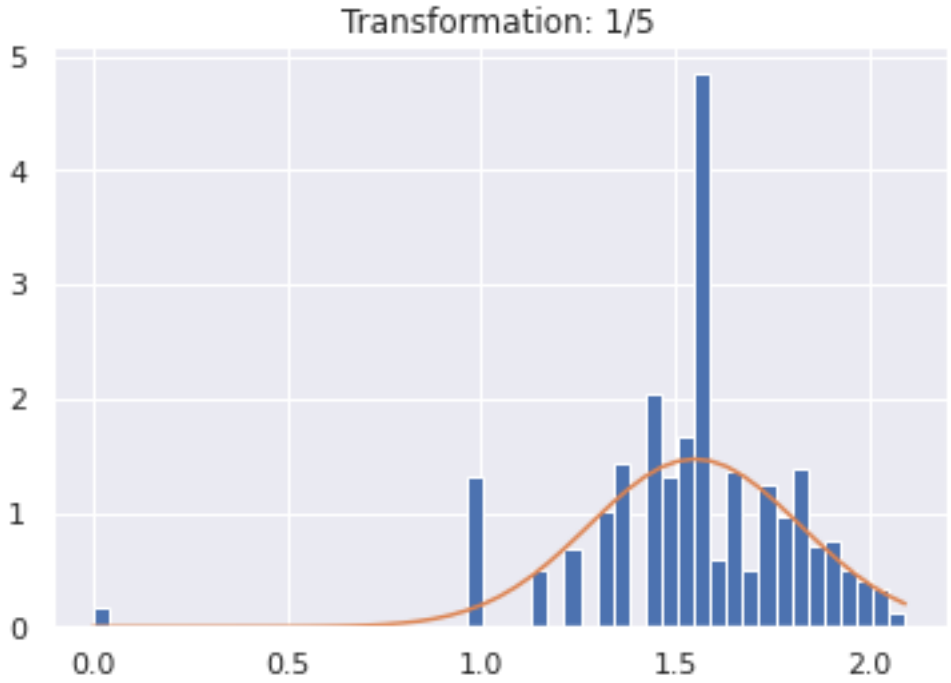

Transformation: 1/3
Sample Quantiles
Theoretical Quantiles
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−3
−2
−1
0
1
2
3

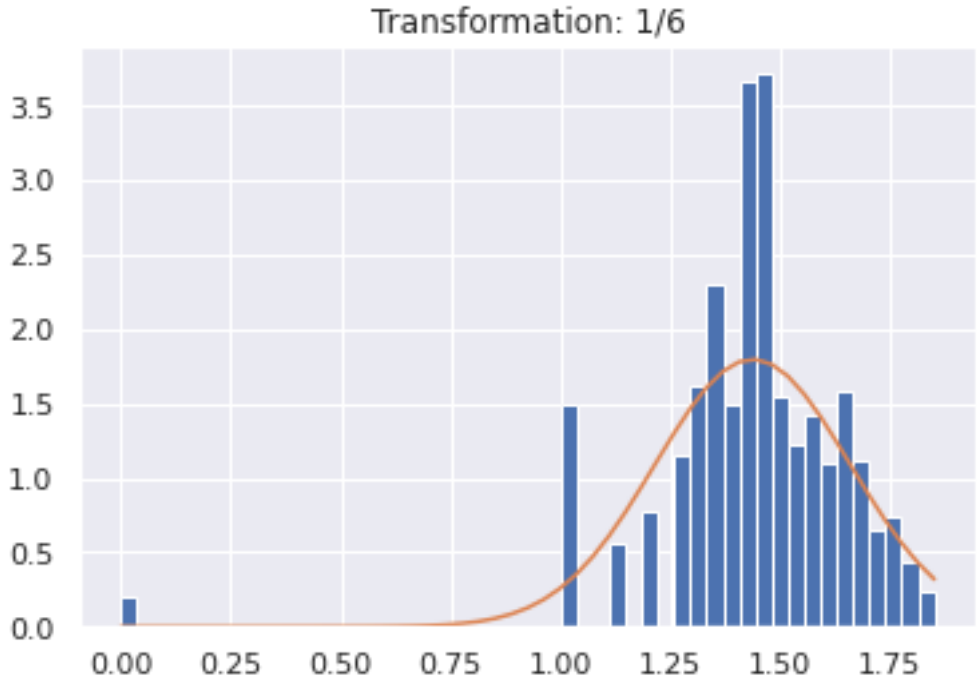

Transformation: 1/6
Sample Quantiles
2.0
1.5
1.0
0.5
0.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Transformation: 1/8
Sample Quantiles
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Transformation: 1/2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6


Transformation: 1/3
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5


Transformation: 1/5
5
4
3
2
1
0
0.0
0.5
1.0
1.5
2.0


Transformation: 1/6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75