Analysis of Employee Attrition and its Various Performance Metrics

Gagan G R
CSE Department
PES University
Bangalore, India
gagan2001j@gmail.com

Kshitij Prit Gopali
CSE Department
PES University
Bangalore, India
gopalikshitij@gmail.com

Navya Eedula
CSE Department
PES University
Bangalore, India
navvu003@gmail.com

Sarthak Deva
CSE Department
PES University
Bangalore, India
sarthak2002deva@gmail.com

***Abstract*— Employee attrition is a prevalent problem in today's globally competitive world. Although it is a common issue, it must be managed effectively by companies. A company's worth and success are determined by how well the employees' benefits, job satisfaction, loyalty, and trust the employees and employers have in each other. We aim to build a model that can help determine the causes and likeliness of employees to attrite from a company. In doing so, the HR manager can make quicker decisions to cater to the needs of the employees and retain them in the company. We wish to implement various models – logistic regression, XGBoost, SVM and KNN model. Our primary goal is to gain insights on the HR attrition scenario amongst employees belonging to different age groups, genders, departments and educational backgrounds and determine the leading factors affecting attrition.**

Keywords*— *HR Attrition, Key Performance Indicators, Employee Attrition, Logistic Regression, KNN, XGBoost, SVM

I. INTRODUCTION

Employee attrition refers to a deliberate, yet gradual reduction in the number of employees working in the company. This can happen due to employees retiring or resigning, but not being replaced. HR attrition can be voluntary, for example, employees leave the company in search of better opportunities, job transfers, or retirements. It can also be involuntary that includes termination of the job, dismissal of underperforming employees, death of an associate, etc.

The performance, success, and reputation of a company depends on how permanent and promising their workforce is *[2]*. However, in the current global scenario, HR attrition is imminent and inevitable. Often, companies wish to retain their most experienced and high-yielding employees. They may provide them higher income, better environment, salary hikes, and other employee benefits to ensure their retainment in the company.

Nevertheless, high attrition rates can become a serious issue. According to Harvard Business Review, attrition is estimated to incur a loss of $27bn dollars to the US economy alone *[7]*. Losing experienced and skillful workers can affect a company's capital, prominence to decrease performance, and eventually lead to dissatisfaction from their customers. Replacing these employees is often a a cumbersome process with not very fruitful results; hiring replacements burden the company with high costs of interviewing, hiring, and training new employees *[4]*.

It has been estimated that about 80% of employee turnover is due to bad decisions made by recruiters while hiring *[7]*. Often, a recruit is assigned a role he is unable to fulfill or his ideologies simply do not align with the company's work culture, thus stimulating the employee to attrite.

The HR department must be proficient in analyzing the 'Key Performance Indicators' of their employees. KPIs act as pre-defined goals for HR managers to help understand and evaluate the employees better. This will result in the growth of both the employees and the company itself. Various KPIs may include - overtime hours, male to female ratio, talent rating, absenteeism at work, employee productivity, employee turn-over rate, etc.

II. REVIEW OF LITERATURE

Study *[1]* establishes that there were both financial (like monthly salary) and non-financial (like job satisfaction, workspace) that influenced employee attrition. Employees prefer to have a transparent office environment and clearly defined goals to make actionable commitments. The authors found that there is no single solution for employee attrition and is specific for every enterprise. Six hypothesis tests were conducted to find significant correlations and dependencies between the variables educational qualifications, motivational factors, age, and work experience with attrition. This was implemented using statistical approaches and plotting bar charts for the corresponding categorical data. However, statistical charts have no room for mathematical models and analysis.

Study *[2]* concluded that though employee salaries might seem like a dominant factor that influences attrition rates,

key factors include job satisfaction. The study also concludes that by hiring employees that live in the vicinity of the company, they are less likely to attrite due to family reasons. The study devised various methodologies like - correlation analysis, t-test, chi-square, factor analysis, one-way ANOVA, and multiple regression. Hypothesis tests were conducted based on the explanatory variables – age, gender, department, and seeking a new job. However, there were a few limitations and assumptions made in the study. This study was (a) conducted on a very small sample of 100 (b) depending on the willingness of the employees to take the survey. These factors might have led to bias or inaccuracy in the results.

Study *[3]* focuses on factors that lead to job dissatisfaction amongst the employees which in turn leads to high attrition rates in an enterprise. The methodologies used in this study include calculating descriptive statistics such as mean, performing correlation, and chi-squared tests for hypothesis testing. The following conclusions were drawn from this study: 3 R's – Respect, Rewards, and Recognition were found to be the most influential explanatory variables in the study. A high correlation was found between these 3 R's and lower attrition rates. However, the scope of this study is limited to medium-scale industries in Hampapuram, Ananthapuramu, and was conducted over a short duration. To maintain confidentiality, the managers did not reveal all the information about their employees; this might have led to some bias.

Paper *[4]* also aimed at developing a predictive model and employed machine learning models like SVM with several kernel functions, random forest, and K-nearest neighbor (KNN) on the original dataset which was imbalanced to gain insights on the factors determining attrition. Attempts were made to balance the dataset by both undersampling and oversampling to check if it increased the accuracy of the model. On the imbalanced dataset, the highest accuracy was found using the SVM model with a quadratic kernel function. This accuracy was 0.50. Accuracy was further improved by balancing the dataset using oversampling. The highest accuracy was obtained when feature selection was induced. The F1 score was 0.90 using 12 selected features and 0.92 using only the top two features.

Paper *[5]* uses the same dataset that we used for our study. Various models like Support Vector Machines (SVM) classification, Linear Support Vector Machines (LSVM), Logistic Regression classifier, K-nearest neighbors (K-NN) Gaussian Naive Bayes, Naive Bayes classifier for multivariate Bernoulli models, Decision tree classifier, random forest classifier have been used. This study conducted an elaborate and extensive EDA to understand the dataset and its various parameters and used a training dataset of 70% for training and the remaining 30% was used for testing and validation. Methods like K-fold cross-validation were applied to validate the dataset. The model that gave the best results was the Gaussian Naïve Bayes model which gave the highest true positive rate of 75% and the lowest false positive rate of 4.5%. This study was focused on lowering the false positive rates (employees who are not likely to leave the company but are classified as so) and improving recall (employees who are likely to leave the company but are not detected by the predictive model) of HR attrition models.

Study *[6]* uniquely approaches the problem statement at hand by attempting to run a logistic regression model to determine key factors that influence attrition rates. The study was conducted across 4000 employees during the course of a year. A logistic regression model was employed on historic data, VIF was implemented to keep track of multicollinearity amongst the predictor variables and a confusion matrix was used to summarize the results. The model had 75% accuracy, 73% recall, 75% specificity and determined 11 key factors that influenced attrition. Interestingly, the study found that employees who were married were more likely to attrite due to family reasons than those who were single or divorced. Attrition was also high amongst junior employees who had fewer working years and less work experience in the company.

III. EXPERIMENT TOOL

We use Python and standard Python libraries to perform data preprocessing, exploratory data analysis, correlation analysis, outlier analysis, data visualization, model building, and model evaluation.

IV. METHODOLOGY AND PROPOSED SOLUTION

We wish to develop a model that is capable of accurately determining:

i) How likely an employee is to attrite from the company (either voluntarily or involuntarily).

ii) What key predictor variables influence employee attrition.

The methodology that we propose to implement is comparable to the previous solutions of this problem statement. However, we are striving towards building a model(s) that will result in better accuracy. The solution we are aiming to arrive at will consist of the following features:

i) Our dataset is highly dimensional, consisting of 35 explanatory variables. We aim to find (a) correlated variables (b) variables that do not have much influence on the target value and in turn apply dimensionality reduction (either PCA or other methods) to avoid the curse of dimensionality and increase the accuracy of the model.

ii) None of the papers we found mentioned the effect of outliers in the dataset. We wish to perform an outlier analysis and understand the effect on attrition and look for any other factors that might have led to this anomaly.

iii) We propose to build different models – logistic regression SVM model, XGBoost, and a KNN model We will perform hyper-parameter tuning and analyze which model would give the most accurate result for the dataset included.

iv)

A. *Assumptions*

(i) Our dataset consists of a fairly large sample of 1470 employees from a reliable and credible data source. This eliminates any bias that might have been induced due to the small size of the dataset (as in *[2]*) or convenient sampling methods (as conducted in *[2]* and *[3]*).

(ii) The dataset has samples from employees in different types of industries, from different educational fields, and different levels of education. This also eliminates a bias factor (study *[3]* was conducted only on medium-scale industrial employees).

Since the dataset is sufficiently large, we use 10% of the data for evaluating and testing the model. The remaining data is split into train and test as 80% and 20% respectively. Since all the records are independent of each other, we deemed this to be a suitable train-test split.

B. *Data Preprocessing*

The dataset consists of 1470 records and 35 attributes with no missing values. It was suitably cleaned and visualized to train the models. The following were the results of the exploratory data analysis:

i) This dataset had 35 features out of which three explanatory variables – employee count, over18, and standard hours were meaningless and irrelevant to the study. All the employees in this dataset are above 18 years of age and work for a minimum of 60 hours a week. These columns were hence dropped to reduce dimensionality.

ii) The categorical data was suitably encoded using binary encoding and label encoding. It was first considered to use one-hot encoding for certain nominal data, however, it was realized that this would further increase the dimensionality of the data by introducing sparse columns.

iii) Bivariate analysis was performed to observe the correlation between the variables. Pearson's correlation, spearman correlation, phi coefficient, and point bi-serial correlation were used between all the different types of variables and it was observed that the features – years at the company, total working years, years since promotion, and years with the current manager are all highly correlated to each other. Monthly income is also correlated to total working years. The rest of the features are mostly independent of each other.

C. *Outlier Analysis*

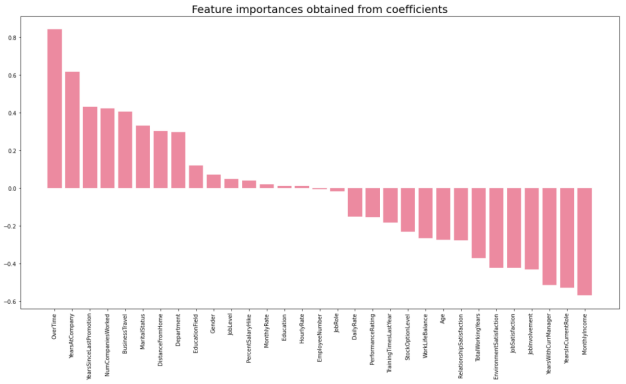
Outliers in the data were detected using box plots and interquartile range to analyze the impact of the outliers on attrition rate and the following insights were drawn:

i) If the monthly salary lies beyond the 75 percentile, then there is only a 4.38% chance that the employee will attrite.

ii) Therefore, if the total working years of an employee lies beyond the 75 percentile, then there is a 7.9% chance that the employee will attrite.

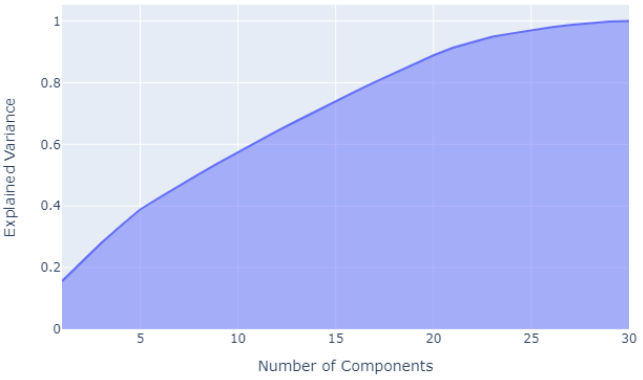
D. *Feature extraction and dimensionality reduction*

The dataset after pre-processing consists of 30 explanatory variables and 1 target value. We attempted to discover the most influential factors affecting HR attrition. Hence, we modeled a logistic regression model for the training dataset and observed the sign and magnitude of the coefficients obtained in the logistic regression equation. The plot of the same is as follows:



Here, it can be concluded that overtime hours and years at the company are the variables that are most positively influencing attrition rates, i.e.., an employee is most likely to attrite if the company makes him/her work overtime. Conversely, factors like monthly income and years in the current role most negatively influence attrition rates, i.e.., an employee that is paid well is least likely to leave the company. These insights align with those we obtained from outlier analysis.

To overcome the curse of dimensionality, we used PCA as a dimensionality reduction technique after scaling the data using StandardScaler.



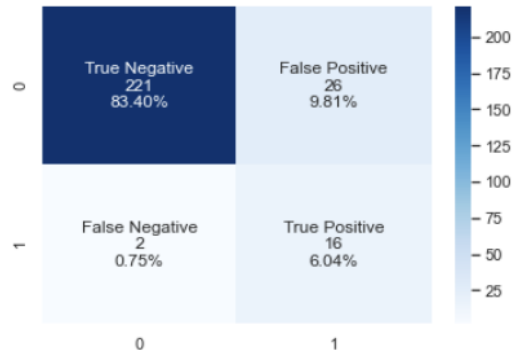
From this graph, it can be observed that 25 features of the dataset retain 97.0% of the data. Hence, the 30 features were reduced to 25 features and used for further model training.

E. *Fitting the models*

A variety of classification models were implemented to build the best model while simultaneously attempting to understand the data better and tweak the model parameters accordingly. Since the range of the data is highly varying in all columns, we have used StandardScaler pre-processing to eliminate the influence of variables with high variance.

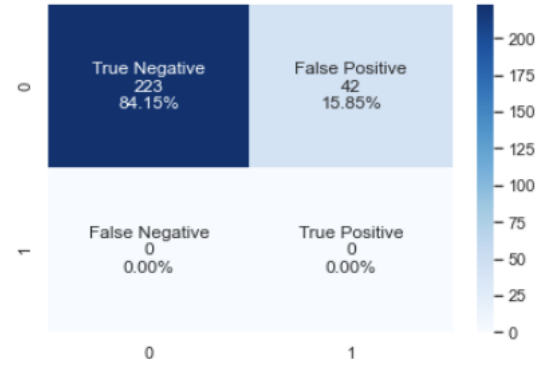
We decided to use logistic regression to predict the outcome of this classification problem. This regression model produces a constant binary output using the predictor variables, unlike a linear or multi-linear regression model. The independent variables are linearly related to log odds.

After performing hyper-parameter tuning, the parameters, C=10000.0, solver='liblinear' trained the model with the highest accuracy. The confusion matrix for the training data is as follows:



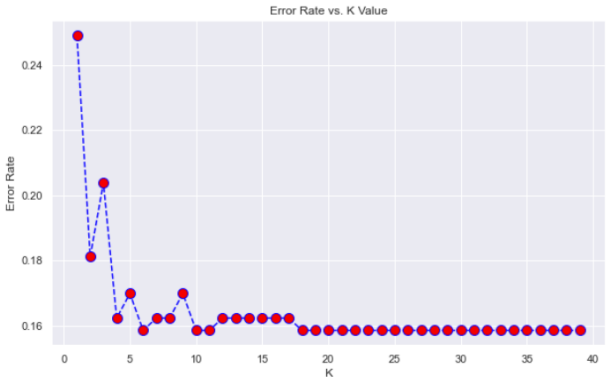
After testing the model on the 10% test set, the accuracy of the model was found to be 89.4% and the F1 score was 94.04%.

We attempted to train the model using XGBoost which is an ensemble model that implements gradient boosted decision trees. On performing hyper-parameter turning, the following parameter gave the least RMSE of 0.166. The confusion matrix for the training data is as follows.

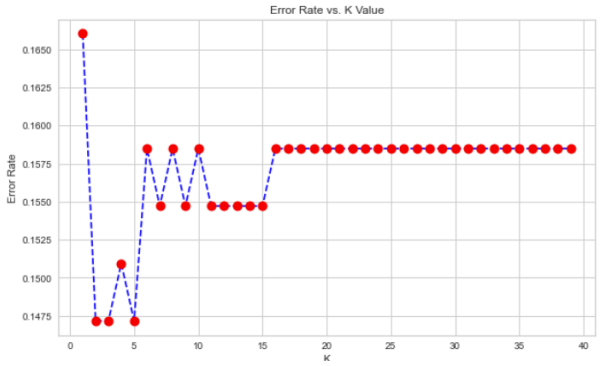


Upon fitting the model on the testing dataset, the F1 score was found to be 93.1%. However, from the confusion matrix, it can be seen that the training model has a 100% recall rate. This indicates the overfitting of the model.

The KNN model was then chosen to train the data. Optimal K value was found for the training dataset using the elbow curve method as shown below:



We decided to choose K=11 to train the model and obtained an accuracy of 89.1% and an F1 score of 94.1% for the test dataset.
Since KNN suffers from the curse of dimensionality, the model training process was repeated after performing dimensionality reduction using PCA. The elbow curve plotted for this dataset is as follows:

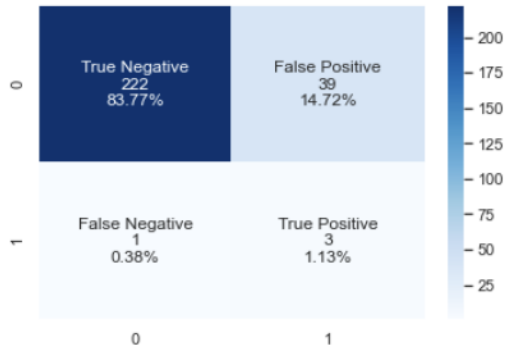


The confusion matrix for the train dataset is as follows:

We achieved an improvement in the accuracy of the model with the accuracy and F1 score of 89.8% and 94.5% respectively for the new test dataset. However, due to this dimensionality reduction, individual features lost their interpretability.

An SVM classifier was implemented as we believed that it would train the model well even for high dimensional data. SVM is very useful for non-linear

classification models as it implicitly maps the input features to higher-dimensional space. Since this is the nature of the dataset, the SVM model proved to give great results. Upon hyper-parameter tuning with the kernels 'rbf' and 'poly', results obtained using the polynomial kernel proved to be more effective. The confusion matrix for the training set is as follows:



The RMSE obtained on the test set was found to be 0.0878 with accuracy, precision, recall, and F1 score of 91.2%, 100%, 90.8%, and 95.2% respectively. The SVM model proved to be the most accurate model of all models trained so far.

V. RESULTS AND CONCLUSIONS

Understanding employee attrition, and analyzing its root causes can help companies retain their most productive employees and this can, in turn, lead to the success of a company. Employee retention must be managed and the The turnover rate must be kept below the target level *[3]*. This
can be done by focusing, investing in, and honing the skills of the HR department.

We have obtained the highest accuracy with the SVM classifier using the polynomial kernel. This accuracy turns out to be 91.2% with an F1 score of 95.2%.

It is becoming increasingly popular in various industries and fields to adopt artificial intelligence and well-trained models such as these in decision-making
activities within a company *[5]*. By using predictive models to analyze employee attrition, quicker and more efficient measures can be put in place by the HR managers to limit employee attrition. The aim is to make management decisions based on "objective data rather than subjective considerations *[5]*".

There are a few factors that influence attrition rates more than the others. The conclusions drawn are stated below:

i) Overworking the employees and making them stay overtime can lead to job dissatisfaction and in turn higher attrition rates.
ii) Employees who have worked in the same company for many years are likely to attrite, or more specifically retire from their job. Many companies inevitably lose experienced and skilled employees due to retirement.
iii) Paying employees high working wages is more likely to build their loyalty to the company and hence, they are more likely to stay in their jobs.
iv) Employees who tend to work for many years in the same job role are most likely satisfied with their job and are less likely to attrite.
v) Factors like age, educational field, department, and job role did not seem to have much of an impact on employee attrition.

After obtaining satisfactory results, we conclude with suitable solutions that can help retain employees in the company. These solutions are based entirely on the
results we obtain from our data model The key performance indicators (KPI's) that turned out to be most insightful according to our study were – overtime work hours, monthly salary, and overall job satisfaction. HR personnel can analyze these KPIs of each of their employees to make a thoughtful
decision on the following.:

i) Job roles to allot to employees based on their specific skill set.
ii) Ensure a healthy, stress-free working environment for each worker.
iii) Develop loyalty and trust amongst the employees whilst inculcating in them the ideologies of the company.
iv) All in all, build a successful company in the competitive global market.

VI. FUTURE PROSPECTS

The models that were tried out were one of many training models that can be used for this dataset. Another efficient supervised model that could be devised could be the decision tree classifier model. Since the dimensionality is high, feature extraction and PCA could be used. Effective ensemble models like bagging and boosting techniques or rule pruning could be applied to further improve the accuracy of the model.

VII. ACKNOWLEDGEMENTS

We humbly and sincerely thank Dr. Gowri Srinivas and the teaching assistants of the Data Analytics course for guiding us through this project. This was indeed a learning opportunity and a thoroughly enjoyable experience. The results we gained were insightful and motivated us to take up projects like these and participate in Kaggle competitions in our future endeavors. Through this journey, we also learned about the attrition scenario in companies and we truly believe that this knowledge would prove to be useful after we graduate. We would also like to thank the Computer Science department at PES University for motivating us to conduct this research while

equipping us with the necessary problem-solving knowledge.

VIII. REFERENCES

[1] Talapatra, Pradip & Rungta, Saket & Anne, Jagadeesh. (2016). EMPLOYEE ATTRITION AND STRATEGIC RETENTION CHALLENGES IN INDIAN MANUFACTURING INDUSTRIES: A CASE STUDY. VSRD International Journal of Business and Management Research. VI. 251-262.

[2] Lavanya, B. Latha. "A Study on Employee Attrition: Inevitable yet Manageable." International Journal of Business and Management Invention 6.9 (2017): 38-50.

[3] Silpa, N. "A study on reasons of attrition and strategies for employee retention." International Journal of Engineering Research and Applications 1.5 (2015): 59-62.K. Elissa, "Title of paper if known," unpublished.

[4] S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," *2018 International Conference on Innovations in Information Technology (IIT)*, 2018, pp. 93-98, doi: 10.1109/INNOVATIONS.2018.8605976.

[5] Fallucchi, F.; Coladangelo, M.; Giuliano, R.; William De Luca, E. Predicting Employee Attrition Using Machine Learning Techniques. *Computers* **2020**, *9*, 86.

[6] Setiawan, I., et al. "HR analytics: Employee attrition analysis using logistic regression." *IOP Conference Series: Materials Science and Engineering*. Vol. 830. No. 3. IOP Publishing, 2020.

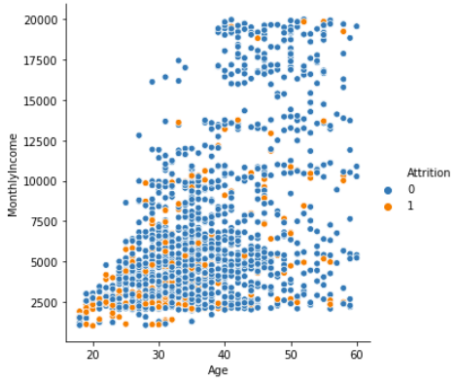
[7] Marsden T. What is the true cost of attrition?. Strategic HR Review. 2016 Aug 8.

[8] Karumuri V, Singareddi S. Employee attrition and retention: A theoretical perspective. Asia Pacific Journal of Research Vol: I Issue XIII. 2014 Jan.

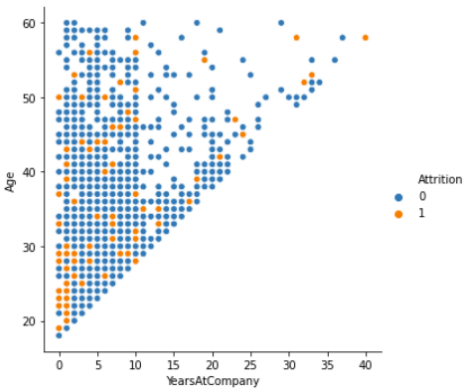
IX. APPENDIX

A. *Insights from EDA:*

i) The scatter plot shows the correlation of age with monthly income and also gives an insight into the attrition rates concerning these features. We can conclude from this that as employees get older, they get better wages in the company and are hence less likely to attrite.



ii) This scatter plot concludes that as employees grow in the same company for many years, they are less likely to attrite and instead stay loyal to the company. A few outliers here could be if the employees retire.

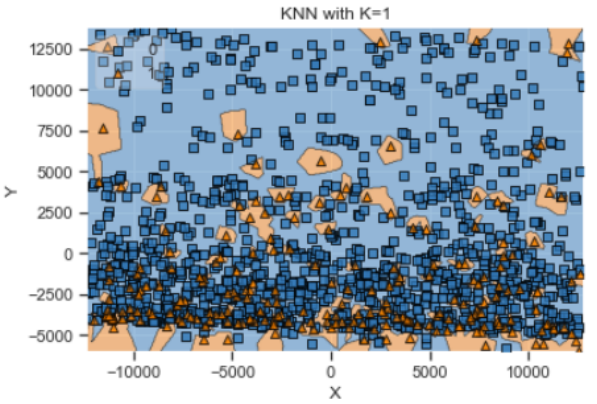


B. *Hyper-parameters for XG Boost model:*

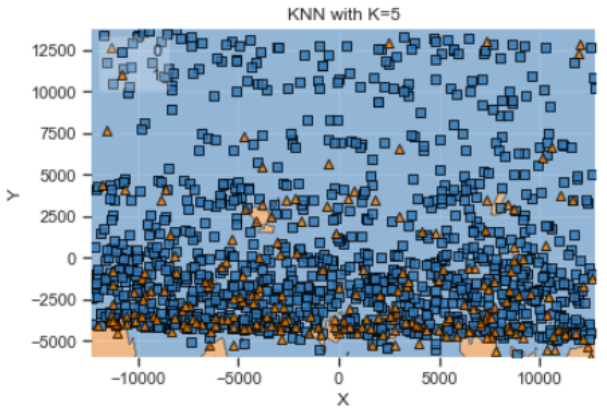
{'colsample_bytree': 0.6536555651994, 'gamma': 5.775981335020807, 'max_depth': 17.0, 'min_child_weight': 4.0, 'reg_alpha': 100.0, 'reg_lambda': 0.8847852026582983}

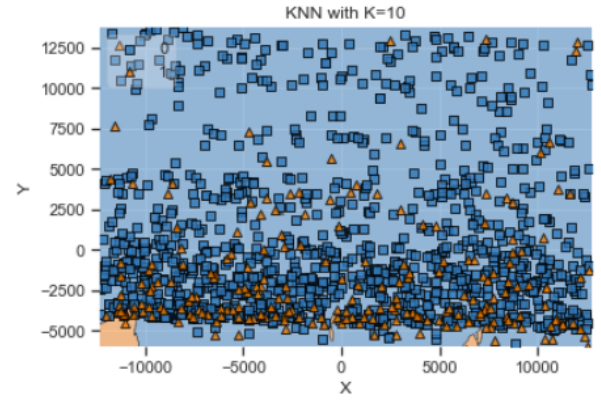
C. *Attempting to visualize the KNN model*

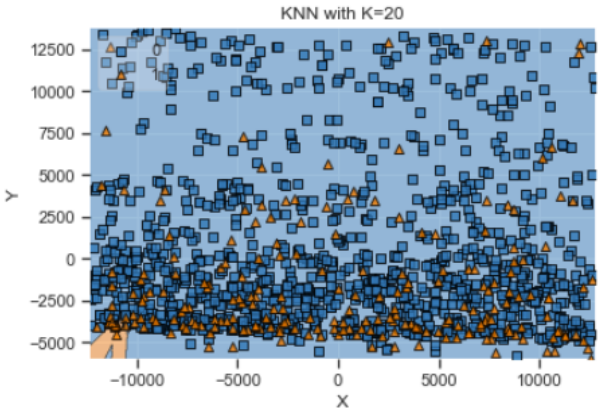
These were the graphs obtained from the visualization of the KN model. However, since our dataset is highly dimensional, it cannot be accurately visualized.



This graph, for K=1, is highly overfitting and hence inaccurate. As the K value increases, the decision boundary becomes smoother as follows:







It is very difficult to obtain visualisations for high dimensional data without reducing the data to two or three dimensions.

D. What we did differently from other papers in the same field:

i) Use statistical and mathematical analysis and models that conclude the dataset that is truly in line with real-life attrition scenarios at companies.

ii) Obtained a high F1 score of 95.2% with the SVM model.

iii) Eliminated all possible biases from the dataset and effective pre-processing to obtain all values in the same range.

iv) Successfully implemented PCA dimensionality reduction technique to avoid the curse of dimensionality and improve model performance.