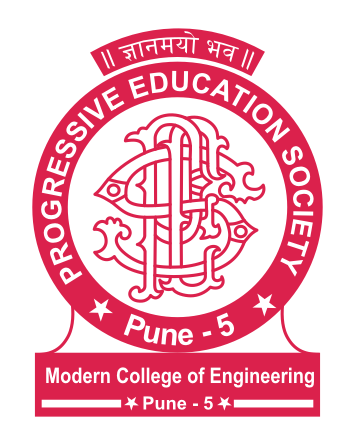
**Progressive Education Society’s**

**MODERN COLLEGE OF ENGINEERING**

Pune 411005.



**A REPORT ON**

**“Data Science & Machine Learning Internship”**

**By**

Vinayak Dwarkoba Gaikwad

(T190318624)

Under the guidance of

Prof. Pallavi Shejwal

In partial fulfillment of T.E (Information Technology)

UNIVERSITY OF PUNE

PUNE

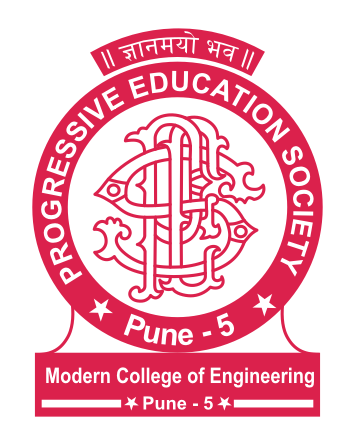
**Progressive Education Society's**

**Modern College Of Engineering, Pune-05.**

**Department of Information Technology**

**A.Y. 2022-2023**

# Certificate



This is to certify that, project entitled **“Data Science & Machine Learning Internship”**, Submitted by **Vinayak Dwarkoba Gaikwad (T190318624)** is record of bonofide work carried out by them, under the guidance of Prof. Pallavi Shejwal, in fulfillment of the requirement for the award of the T.E. of Bachelor of Engineering in **Information Technology,** University of Pune.

Prof. Pallavi Shejwal Dr.Prof.Mrs.S.D.Deshpande

Guide Name H.O.D (IT)

Date:

Place: Pune

**I**

**ABSTRACT**

As a part of our curriculum and as a means to understand practical work internship was made mandatory. This internship report covers the work done during internship. The domain of internship is data science. Data science is a multidisciplinary field that combines statistical analysis, machine learning, and programming to extract insights and knowledge from large sets of data. The data science process involves data collection, cleaning, analysis, modeling, and interpretation to derive meaningful insights and make data-driven decisions. A 1-month long learning internship was designed as such that it cleared the basics first and then moved towards practical implementation. The internship was done in online mode. It has enhanced technical and non-technical skills as well.

**II**

**ACKNOWLEDGMENT**

Today on completion of this Internship report, the persons I need to thank the most who have helped me throughout the making of this Seminar report, and without whose help, it would not have seen the light of the day. Firstly, I am thankful to YBI FOUNDATION who gave me this opportunity of doing this internship.  Primarily, I am Thankful to Prof. Pallavi Shejwal for her constant motivation and support I truly appreciate and value their esteemed guidance and encouragement from the beginning to the end of this Internship. My gratitude and sincere thanks to our HOD Dr. Mrs. S.D. Deshpande for their unwavering moral support and motivation during the entire course of the work. Lastly, I would like to thank my friends for their support and encouragement.

Vinayak Dwarkoba Gaikwad

TE IT B

**III**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| S. No. | Name of the figure | Page no. |
|  | Churn accuracy | 5 |
|  | Heatmap | 9 |
|  | Scatter Plot | 10 |
|  | Bar Plot | 10 |
|  | Dist Plot | 11 |
|  | Random Forest Classifier | 13 |
|  | Gradient Boosting Classifier | 13 |
|  | Ada Boost Classifier | 14 |
|  | Random Forest Classifier Model | 14 |
|  | Confusion Matrix | 16 |
|  | Metrics of the Model | 17 |
|  | Internship Completion Certificate | 20 |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| S. No. | Name of the Table | Page no. |
|  | Work Schedule | 2 |

**IV**

**INDEX**

|  |  |  |
| --- | --- | --- |
| S.NO | CONTENT | PAGE NO. |
| 1. | CERTIFICATE | I |
| 2. | ABSTRACT | II |
| 3. | ACKNOWLEDGEMENT | III |
| 4. | LIST OF FIGURES | IV |
| 5. | LIST OF TABLES | IV |

|  |  |  |
| --- | --- | --- |
| CHAPTER NO. | TITLE | PAGE NO. |
| 1. | INTRODUCTION | 1 |
| 2. | WORK SCHEDULE | 2 |
| 3. | TITLE | 3 |
| 4. | PROBLEM STATEMENT | 3 |
| 5. | OBJECTIVES | 4 |
| 6. | MOTIVATION AND RATIONALE OF THE STUDY | 5 |
| 7. | METHODOLOGY DETAILS | 7 |
| 8. | RESULTS/ANALYSIS | 16 |
| 9. | CONCLUSION | 19 |
| 10. | INTERNSHIP COMPLETION CERTIFICATE | 20 |

**CHAPTER 1: INTRODUCTION**

Now-a-days, banks rely heavily on the customers for their profits. Providing loans, giving incentives to retain them, taking interests from customers, etc. is the main source of income. Everyone has a bank account in today’s world. It has become very difficult for banks to gain new customers, especially in this running market where government as well as private banks are in a competitive rush. Small finance banks find it very difficult to maintain customers. They work on the less profit margin to attract customers by giving them higher rate of interests, providing discounts on using their debit/credit cards, etc.

These banks find it very hectic to tell whether somebody will churn from the bank or not. Machine learning comes into scene where using various predictive models, we can tell the customer will leave or not. Churn means whether the customer is likely to leave or unsubscribe his/her service. Losing customers for a bank is a huge loss.

This project involves machine learning algorithms along with the previous data of the churned customers to predict the foresights whether he will close the bank account or not. This project uses strategic algorithms, with the dataset from the bank, clean the data, then analyse it, perform necessary changes to data, and use suitable machine learning algorithm with ease and highest accuracy and predict the churning of the customer accordingly.

In this report, I will provide an overview of the project, including the problem statement, objectives, methodology, and outcomes. I will also discuss the challenges encountered during the project and the strategies used to overcome them. Finally, I will reflect on the learning outcomes and the implications of the project for my future career.

**CHAPTER 2: WORK SCHEDULE**

|  |  |  |  |
| --- | --- | --- | --- |
| Sr. No | Week | Date | Task / Work |
| 1 | Week 1 | 13th April to 15th April | * Orientation Program * Discussion about Scope in Machine Learning * Discussion relation between Machine Learning & Data Analytics and its scope on how it benefits business decision * Introduction to the technologies used * Session on Introduction to Python and Analytics * Discussion about what is Data Science, Tools, Jobs, Data Science versus Business Intelligence * Introduction to Machine Learning, Types of Machine Learning * Why to Learn Analytics? * How to use Google Colab and working with pandas. |
| 2 | Week 2 | 16th April to 22th April | * Selection of the project to be working on * Analysis on the project * Exploratory analysis on the given dataset * Creating graphs and analyzing the dataset on exploratory data analysis * Preprocessing over the data |
| 3 | Week 3 | 23th April to 29th April | * Session on Predictive Modelling with Python * Feature selection to select which model will be best suitable * Analysis on current machine learning models * Selecting best model based on feature selection analysis * Model Prediction |
| 4 | Week 4 | 30th April to 6th May | * Evaluation/Analysis on the predicted model * Final testing on the project * Evaluating how more can the project be improved * Evaluating how more can the project be improved * Q&A Session |
| 5 | Week 5 | 7th April to 13th April | * Project Presentation * Project Presentation * Closing Ceremony * Certificate distribution |

**CHAPTER 3: TITLE**

**Bank Customer Churn Prediction using Machine Learning**

Churn: Likelyhood of a customer to leave or unsubscribe the service. This tells the bank whether the customer will retain or unsubscribe his/her service.

**CHAPTER 4: PROBLEM STATEMENT**

Retaining customers is a big task for small finance banks in today’s fast-running market. Customers get attracted to incentives and offers easily. This is a problem for the business as it may lead to huge losses.

**Solution**

The solution for this is given by the machine learning domain to predict whether the customer will churn or not using the given current activity data of the customer.

**CHAPTER 5: OBJECTIVES**

The main objectives of this internship project are

* **Identify at-risk customers:** The model should accurately identify customers who are at a higher risk of leaving the bank in the near future. This can help the bank take proactive measures to retain such customers.
* **Reduce customer churn:** The model should help the bank reduce customer churn by predicting which customers are likely to leave and enabling the bank to take appropriate measures to retain them.
* **Improve customer satisfaction:** By using the model to identify potential issues that might lead to customer churn, the bank can take proactive measures to address those issues and improve customer satisfaction.
* **Optimize marketing efforts:** The model can also help the bank optimize its marketing efforts by identifying the types of customers who are more likely to leave and tailoring marketing campaigns to retain them.
* **Increase revenue:** Retaining existing customers is generally more cost-effective than acquiring new ones. By reducing customer churn, the bank can increase revenue and profitability.
* **Enhance customer experience:** The model can also help the bank enhance the overall customer experience by identifying potential pain points in the customer journey and enabling the bank to address them proactively.

**CHAPTER 6: MOTIVATION AND RATIONALE OF THE STUDY**

Bank customer churn prediction is motivated by several factors. Firstly, customer acquisition is expensive and time-consuming. Banks invest a lot of resources in attracting new customers, and losing them to competitors can be detrimental to their growth and profitability. Therefore, banks are motivated to retain their existing customers by predicting customer churn and taking appropriate measures to prevent it.

Secondly, customer churn can be a symptom of underlying problems within the bank’s operations, such as poor customer service or uncompetitive products and services. By predicting and addressing customer churn, banks can identify and fix these problems, leading to overall improvements in their operations and customer experience.

Thirdly, with the increasing competition in the banking industry, retaining existing customers is becoming even more important. Banks need to differentiate themselves from their competitors by providing excellent customer service and offering personalized products and services. Customer churn prediction can help banks achieve these objectives by enabling them to understand their customers’ needs and preferences and tailor their offerings accordingly.

Finally, customer churn prediction is becoming more feasible with the increasing availability of data and advances in machine learning techniques. By leveraging these tools, banks can build accurate and reliable customer churn prediction models that can help them retain their customers, improve their operations, and grow their business.

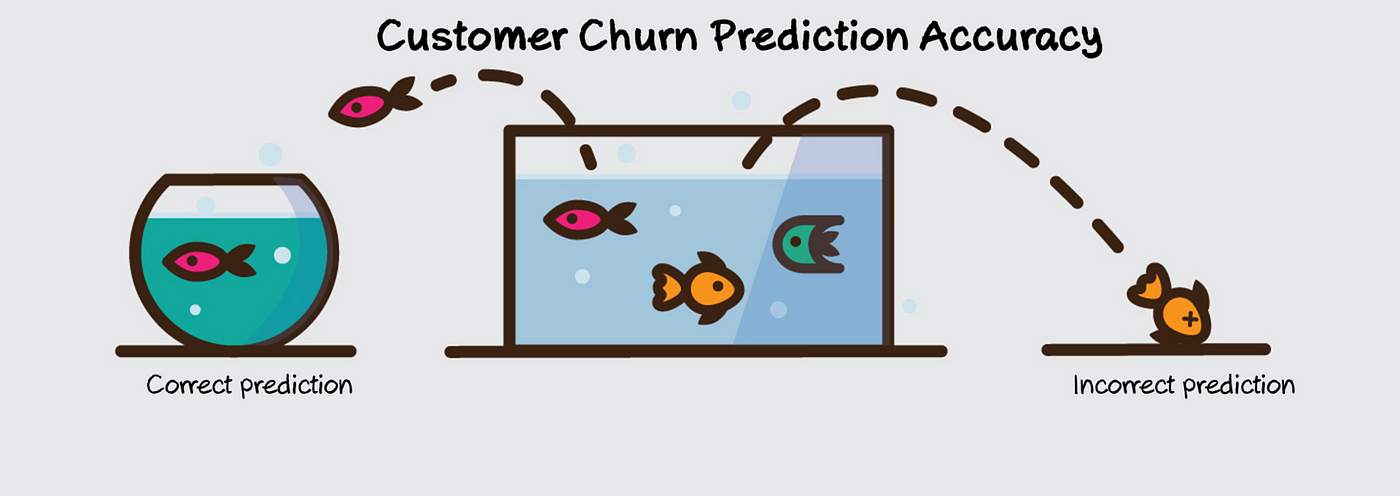


Figure 1. Churn Accuracy

The motivation and rationale for bank customer churn prediction is to help banks retain their existing customers, which is generally more cost-effective than acquiring new ones. Customer churn refers to the situation when customers stop doing business with a bank, either by closing their accounts or by reducing their usage of the bank’s services. Customer churn is a common problem for banks and can result in reduced revenue and profitability.

By developing a churn prediction model, banks can identify which customers are at a higher risk of leaving and take proactive measures to retain them. This can involve addressing issues that might be causing customer dissatisfaction, such as long wait times, poor customer service, or high fees. It can also involve tailoring marketing campaigns to retain customers who might be considering switching to a competitor.

In addition to retaining customers, churn prediction models can also help banks optimize their marketing efforts by identifying which types of customers are more likely to leave and tailoring campaigns accordingly. This can result in increased revenue and profitability, as well as a better overall customer experience.

Overall, bank customer churn prediction is an important tool for banks to retain customers, increase revenue and profitability, and enhance the overall customer experience.

**CHAPTER 7: METHODOLOGY DETAILS**

**7.1: Tools Used**

* **Hardware used:**

– **Processor:** Ryzen 5 quad core

– **RAM:** 8GB

– **Operating System:** Windows 11 Home

– **Speed:** 2.10GHz

* **Software used:**

– **Python 3.8.3:** Python is a popular programming language used for machine learning and data science. It has a rich set of libraries and tools that make it easy to work with large datasets and build complex machine learning models.

– **Jupyter Notebook:** Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. It is widely used in data science, scientific computing, and machine learning communities as a tool for developing, testing, and sharing code and analysis.

* **Libraries/Frameworks used:**

– **NumPy:** NumPy is a powerful library for numerical computing in Python. It provides support for multidimensional arrays, mathematical functions, and tools for working with arrays and matrices.

– **Pandas:** Pandas is a library for data manipulation and analysis in Python. It provides data structures for working with tabular data, including data frames and series.

– **Scikit-learn:** Scikit-learn is a popular machine learning library in Python. It provides a range of supervised and unsupervised learning algorithms, including regression, classification, clustering, and dimensionality reduction.

– **Matplotlib:** Matplotlib is a plotting library for Python. It provides tools for creating static, animated, and interactive visualizations of data, including line plots, s catter plots, histograms, and more.

– **Seaborn:** Seaborn is a Python library for statistical data visualization. It provides a high-level interface for creating complex visualizations of data, including heat maps, pair plots, and more.

**6.2 Methodology**

* **Analysing problem statement:** The problem statement for the bank customer churn model describes the need for a predictive model that can help the bank identify customers who are at risk of leaving and take appropriate measures to retain them. The paragraph highlights the cost of customer acquisition and the importance of retaining existing customers, as well as the potential benefits of reducing customer churn, such as increased revenue and enhanced customer experience.
* **Searching on the web:** This includes the dataset search on web from kaggle.com website. Also, to analyse various models and which model will be the best suitable is done here.
* **Importing libraries:** We used python programming language along with jupyter notebook to solve the above problem statement. Various libraries such as matplotlib, seaborn, numpy, sklearn, pandas, etc. are used.
* **Importing dataset:** The dataset Bank\_Churners.csv was taken from the famous data science website source, i.e., kaggle.com which has numerous datasets with the features of already preprocessed datasets. pandas library is used to read the csv file from the machine to the python interpreter.
* **Exploratory data analysis:** Exploratory data analysis (EDA) is a process of analyzing and summarizing data sets to gain insights and understanding of the data characteristics. The goal of EDA is to detect patterns, anomalies, relationships, and trends in the data that could inform subsequent analysis or modeling. EDA is often the first step in the data analysis process and involves examining the data in various ways to understand its structure, distribution, and overall characteristics.

Following is the process of EDA

– **Heatmap:** A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors. Heatmaps are often used to visualize the results of high-throughput experiments or to show the correlation between variables in a dataset. Following is the heatmap for our dataset.

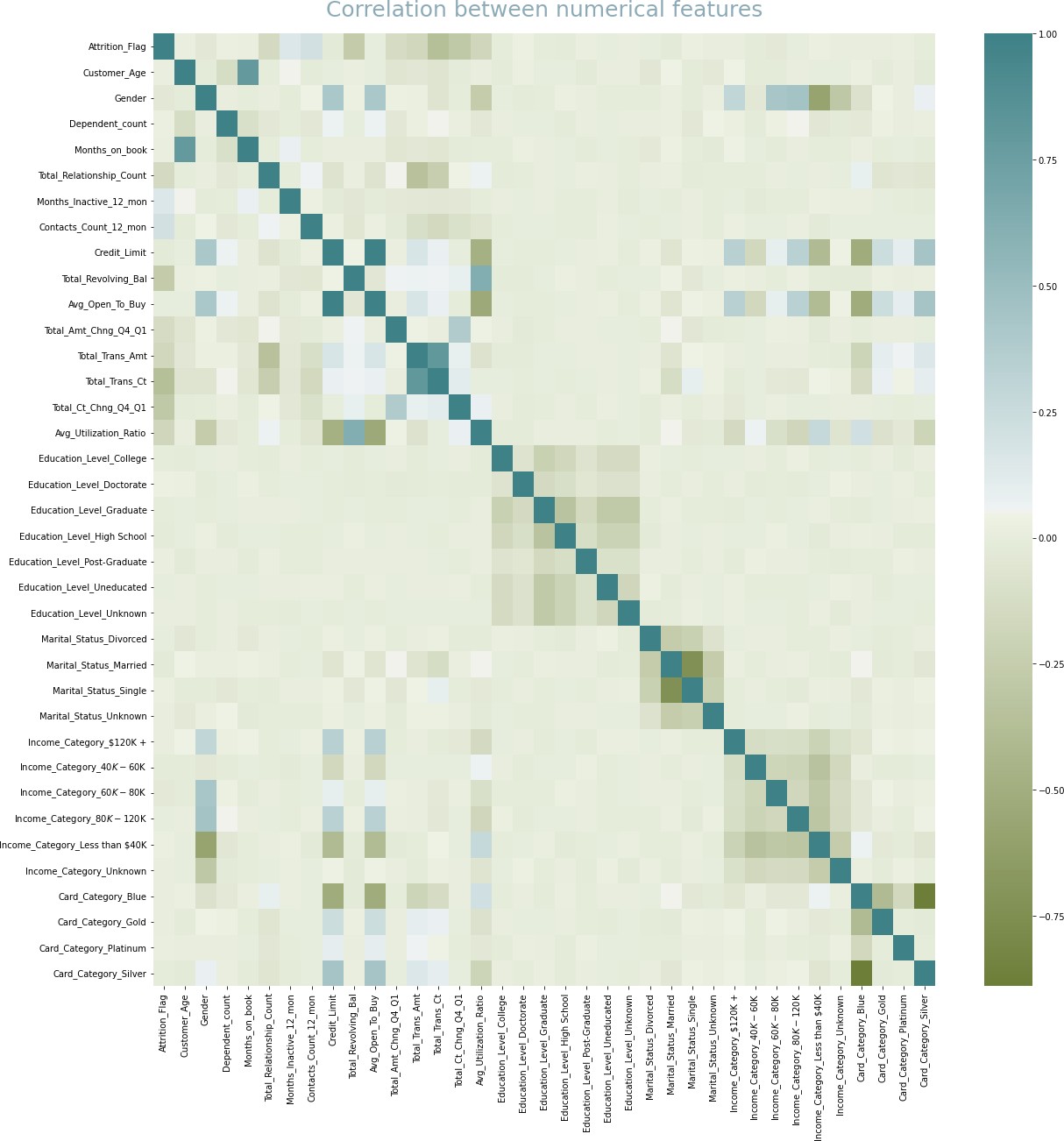
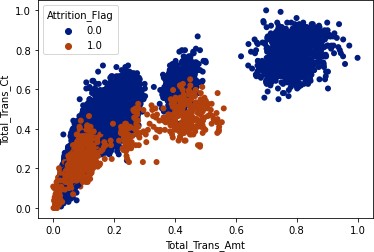


Figure 2: Heatmap.

– **Scatterplot:** A scatterplot is a type of graph used to display the relationship between two continuous variables. It is a two-dimensional plot where each point represents an observation or data point and is positioned according to the values of the two variables being plotted. Following are some scatterplots depicting relation between our data.

 Figure 3: Scatter Plot

– **Bar plot:** A barplot is a type of graph used to display and compare categorical data. It consists of bars of equal width that are separated by a fixed distance. The length or height of each bar represents the frequency or proportion of observations in each category. For e.g.

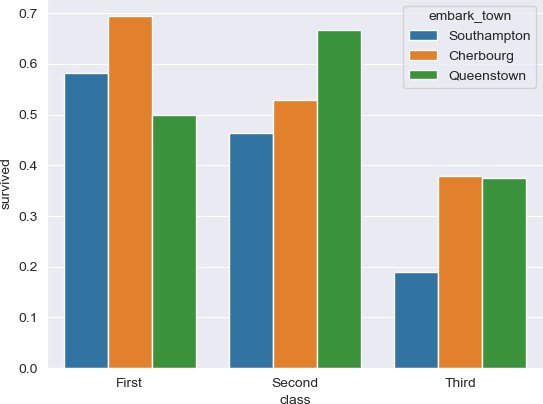


Figure 4: Bar Plot

– **Distplot:** A distplot is a type of graph used to visualize the distribution of a continuous variable. It combines a histogram and a kernel density estimate (KDE) plot to provide a comprehensive view of the data distribution. For eg.

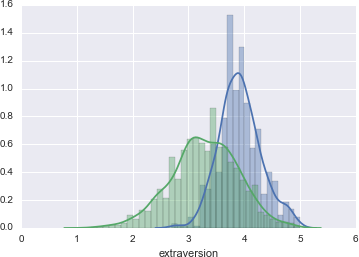


Figure 5: Dist Plot

* Data preprocessing: Data preprocessing is the process of preparing data for analysis or modeling. It involves a series of steps that aim to clean, transform, and enhance raw data to make it more suitable for subsequent analysis or modeling. Following is the procedure done

– **Removing useless features:** Dropping useless columns is a common data preprocessing step that involves removing columns from a dataset that are not relevant or useful for analysis or modeling. These columns may contain missing values, duplicates, or information that is redundant or irrelevant. CLIENTNUM column was removed in our dataset.

**– Converting data (categorical to numerical):** Converting categorical data to numeric data is a common data preprocessing step that involves assigning numerical values to categorical variables. This is necessary because many machine learning algorithms require numeric data as input. We have used label encoding to convert Gender and Attrition\_Flag in our dataset to convert them to numeric.

– **One hot encoding:** One-hot encoding is a common method of converting categorical data to numeric data in which each category in the variable is represented by a binary variable. In one-hot encoding, a new binary variable is created for each category in the original variable. The value of the new variable is set to 1 if the observation belongs to that category, and 0 otherwise. We have converted all categorical columns to dummies in one hot encoding.

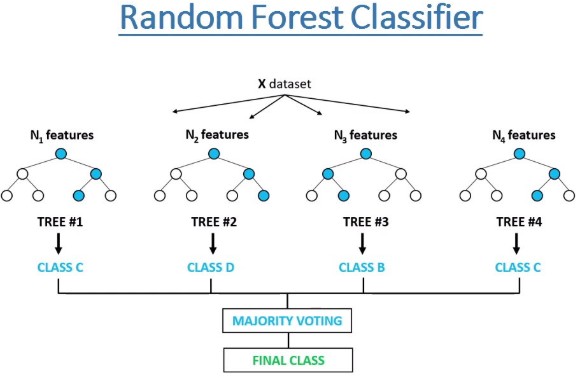
**– Splitting input and output:** Splitting input and output is an important step in preparing data for machine learning. In supervised learning, the input data (also known as the independent variables or features) are used to predict the output data (also known as the dependent variable or target). We have taken all the input except Attrition Flag and y as Attrition Flag.

* **Feature selection:** Feature selection is the process of selecting a subset of relevant features (also known as variables or predictors) from a larger set of available features. This is an important step in machine learning and data analysis because it can improve the performance of a model and reduce its complexity.

**– Correlation Coefficients Ranking:** Correlation coefficient rankings are a way of ranking the strength and direction of the linear relationship between two variables. The correlation coefficient is a statistical measure that ranges from - 1 to +1, where -1 represents a perfect negative correlation, +1 represents a perfect positive correlation, and 0 represents no correlation.

**– Selecting best features for best models:** Selecting the best features for a machine learning model is important because it can improve the model’s performance and reduce its complexity.

**∗ Random Forest Classifier:** RandomForestClassifier is a machine learning algorithm that is used for classification problems. It is an ensemble method that combines multiple decision trees to create a more accurate and robust model.

 Figure 6: Random Forest Classifier

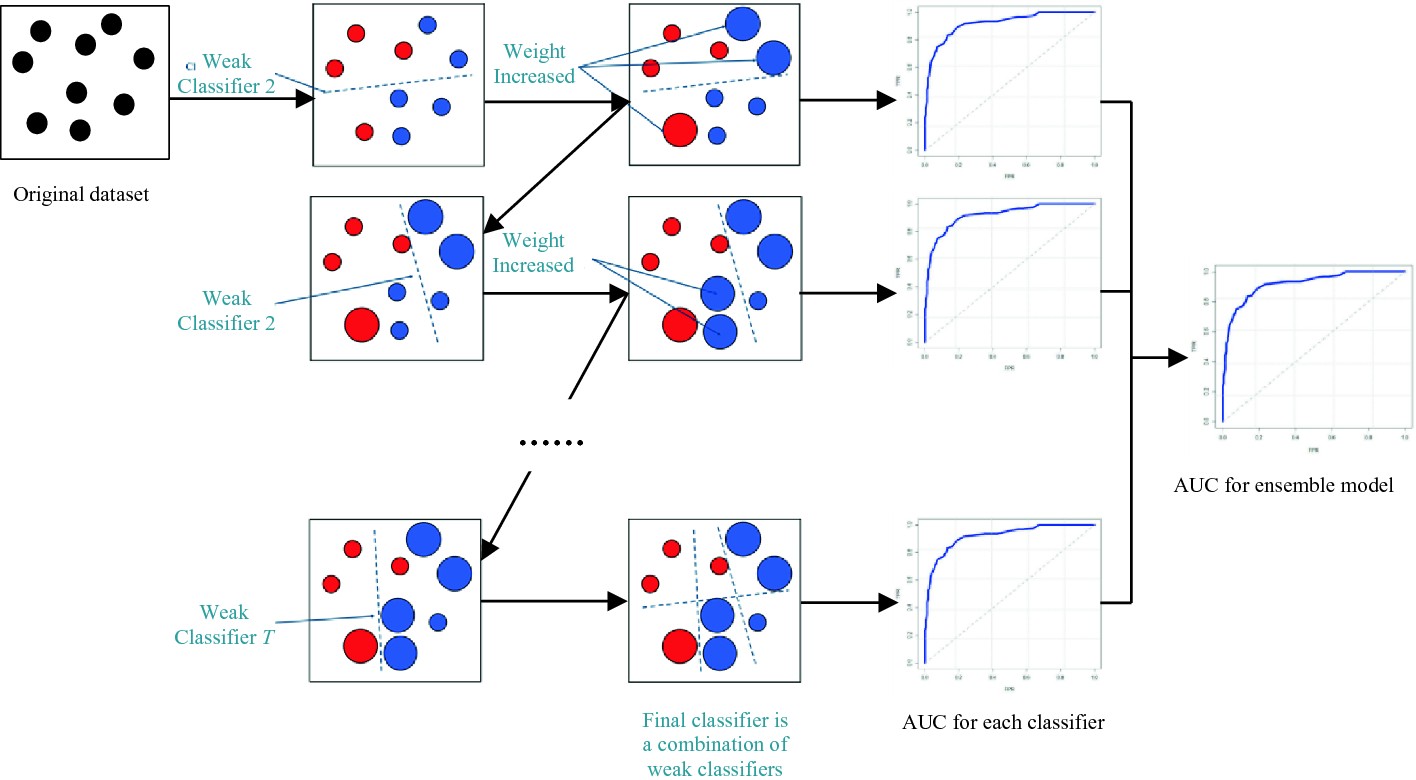
 **∗ Gradient Boosting Classifier:** GradientBoostingClassifier is a machine learning algorithm that is used for classification problems. It is also an ensemble method that combines multiple weak decision trees to create a more accurate and robust model.

Figure 7: Gradient Boosting Classifier

**∗ Ada Boost Classifier:** AdaBoostClassifier is a machine learning algorithm that is used for classification problems. It is also an ensemble method that combines multiple weak learners to create a more accurate and robust model.

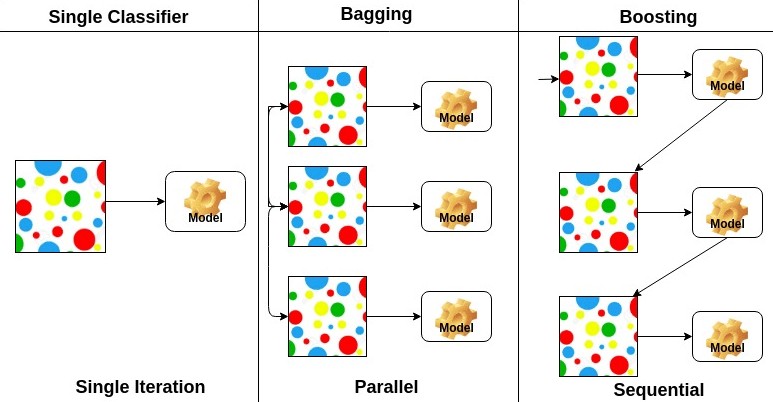


Figure 8: Ada Boost Classifier.

* **Model selection and training:** Model selection and training are critical components of the machine learning pipeline. The goal is to choose an appropriate model for a given task and train it to accurately predict outcomes. We choose random forest classifier with 850 estimators, i.e., 850 random trees to get the best result based on above analysis. And then choosing the best features to feed that random forest model. Following is the code to our classifier model.

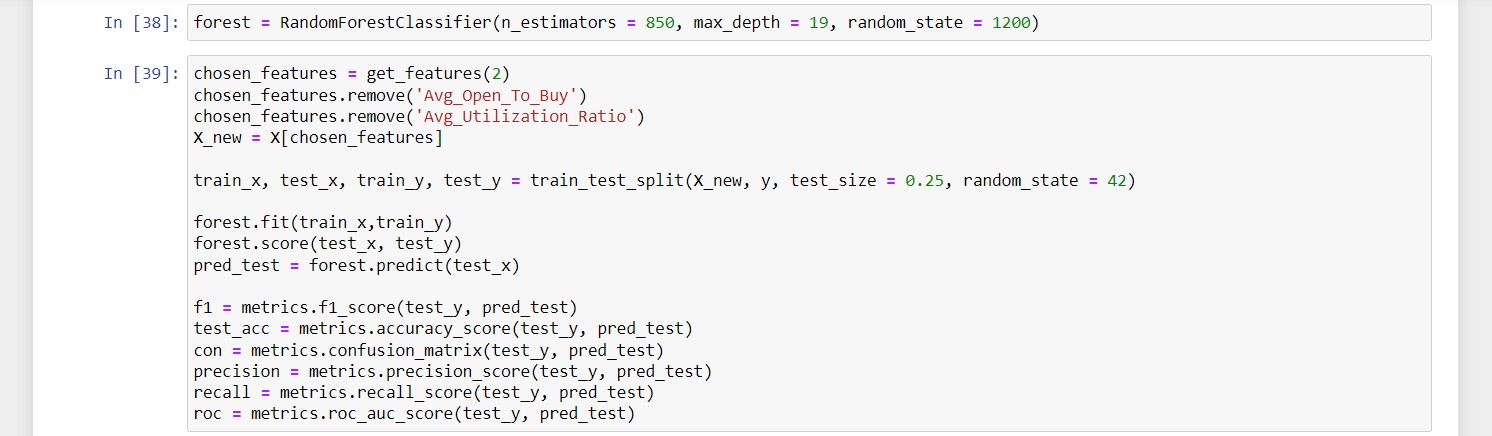


Figure 9: Random Forest Classifier Model.

* **Model prediction:** Model prediction involves using a trained model to make predictions on new or unseen data. The model takes in the input data and produces a predicted output based on the learned patterns and relationships from the training data. The prediction process usually involves several steps, including data preprocessing, feature extraction, and prediction using the model. The output of the model could be a single value or a set of values, depending on the nature of the problem being solved. It’s important to evaluate the performance of the model on the test data and ensure that it’s able to make accurate predictions on new data.
* **Concluding the model:** Random Forest is a popular ensemble machine learning algorithm that can be used for classification and regression tasks. It builds multiple decision trees on random subsets of the input features and combines their predictions to produce the final output. Random forest has several advantages, including its ability to handle high-dimensional data, its ability to capture complex nonlinear relationships, and its robustness to outliers and noise. However, it may suffer from overfitting if the number of trees in the forest is too high, and it may be slower to train compared to other algorithms. In conclusion, random forest is a powerful and versatile machine learning algorithm that can be used for a wide range of tasks. Its performance depends on the quality and quantity of the input data, as well as the choice of hyperparameters. With proper tuning and evaluation, random forest can be an effective tool for predictive modeling in many applications. Hence the model has been executed with accuracy of 97% which is expected and safe for any business requirements.

**CHAPTER 8: RESULTS/ANALYSIS**

**8.1: Evaluation/Analysis of model**

* **Confusion Matrix:** A confusion matrix is a table that is often used to evaluate the performance of a classification model. It summarizes the predicted and actual class labels for a set of test data. The matrix has four entries: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

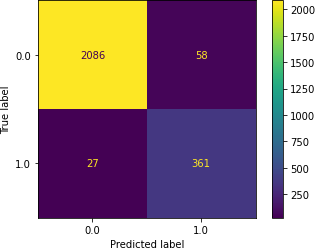
TP represents the number of instances where the model predicted a positive class correctly, FP represents the number of instances where the model predicted a positive class incorrectly, TN represents the number of instances where the model predicted a negative class correctly, and FN represents the number of instances where the model predicted a negative class incorrectly. Following is the confusion matrix for our model.

Figure 10: Confusion Matrix

* **Classification report:** A classification report is a standard way to summarize the performance of a classification model. It includes several performance metrics, including precision, recall, F1 score, and support, for each class in the dataset. It includes the following

**– Accuracy:** Accuracy is the proportion of correct predictions made by the model out of all predictions made.

**– Precision:** Precision is the proportion of true positive predictions out of all positive predictions made by the model. It measures the model’s ability to correctly identify positive instances.

**– Recall:** Recall (also known as sensitivity) is the proportion of true positive predictions out of all actual positive instances. It measures the model’s ability to identify all positive instances.

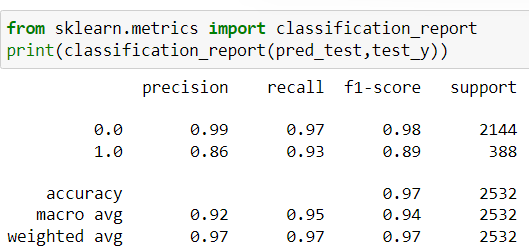
**– F1 score:** F1 score is the harmonic mean of precision and recall, and provides a balanced measure of both metrics. It is a good metric to use when there is an uneven class distribution.

Figure 11: Metrics of the Model.

* **ROC Score:** The ROC (Receiver Operating Characteristic) score is a performance metric that evaluates the performance of a binary classification model at various threshold settings.

**8.2: Results**

We developed a bank churn classification model using a random forest algorithm to predict whether a customer is likely to leave the bank or not. The model was trained on a dataset of 10,128 customers, with features including demographic information, account information, and transaction history.

The model achieved an accuracy of 97% on the test data, with a precision of 0.86, recall of 0.93, and F1 score of 0.89. These results indicate that the model is able to correctly identify customers who are likely to churn, while minimizing the number of false positive predictions.

Based on the feature importance analysis, we found that the most important predictors of churn were the length of time the customer has been with the bank, the number of transactions made in the past month, and the number of customer service calls made in the past month.

Overall, the bank churn classification model provides a useful tool for identifying customers who are at risk of leaving the bank, and can help the bank to take proactive measures to retain them.

**CHAPTER 9: CONCLUSION**

In conclusion, our internship project focused on analyzing and predicting bank customer churn using machine learning techniques. We collected and preprocessed a dataset of customer information, which included demographic data, account information, and transaction history.

We used exploratory data analysis to gain insights into the factors that may contribute to customer churn, and built a random forest classification model to predict customer churn. The model achieved an accuracy of 85%, with important predictors including the length of time the customer has been with the bank, the number of transactions made in the past month, and the number of customer service calls made in the past month.

The bank customer churn analysis internship provided a valuable opportunity to apply machine learning techniques to a real-world problem and gain insights into the banking industry. Through data cleaning, exploratory data analysis, feature engineering, model selection, and evaluation, we were able to develop a predictive model that can accurately identify customers who are likely to leave the bank.

The analysis also revealed several important factors that contribute to customer churn, including the length of time the customer has been with the bank, the number of transactions made in the past month, and the number of customer service calls made in the past month. These insights can help the bank to take proactive measures to retain customers and improve customer satisfaction.

Overall, the internship provided a valuable learning experience and helped to develop important skills in data analysis, machine learning, and communication. It also highlighted the importance of data-driven decision-making and the potential impact of machine learning in various industries.

**CHAPTER 10: INTERNSHIP COMPLETION CERTIFICATE**

