Expresso Customer Churn Prediction

Group Name: AAF Group

Group Members:

First name	Last Name	Student number
Andy	Nguyen	C0891756
Ashok	Kumar	C0894469
Neal	Altares	C0894540

Submission date: ***August 21, 2023***

Contents

Abstract

Churn rate, sometimes known as attrition rate, is a term indicating the likelihood that customers stop doing business with a company after a given period. The higher the churn rate, the more clients stop using products or services from a business. Therefore, churn rate prediction using Artificial Intelligence and Machine Learning algorithms becomes necessary and valuable to identify which customers might be at risk of churning. With the help of these kinds of predictions, businesses can anticipate and optimize their marketing strategies to remain at customer attrition levels under control.

Introduction

In this Churn Prediction report, Expresso, an African telecommunications company, provides customers with airtime and mobile data bundles, shares its data and the desire to predict the likelihood of every customer churning. Expresso puts its focus on two African markets which are Mauritania and Senegal. The ultimate goal of this project's result is to help Expresso have a clear picture of its customers and understand which customers are potentially about to leave.

The reason why we have chosen this topic is that we want to understand the behavioural factors of customers and ultimately forecast as well as develop strategies to retain high-risk customers. This topic is attractive and appealing which could be a first-hand project for us to explore the data in the real-world context.

Methods

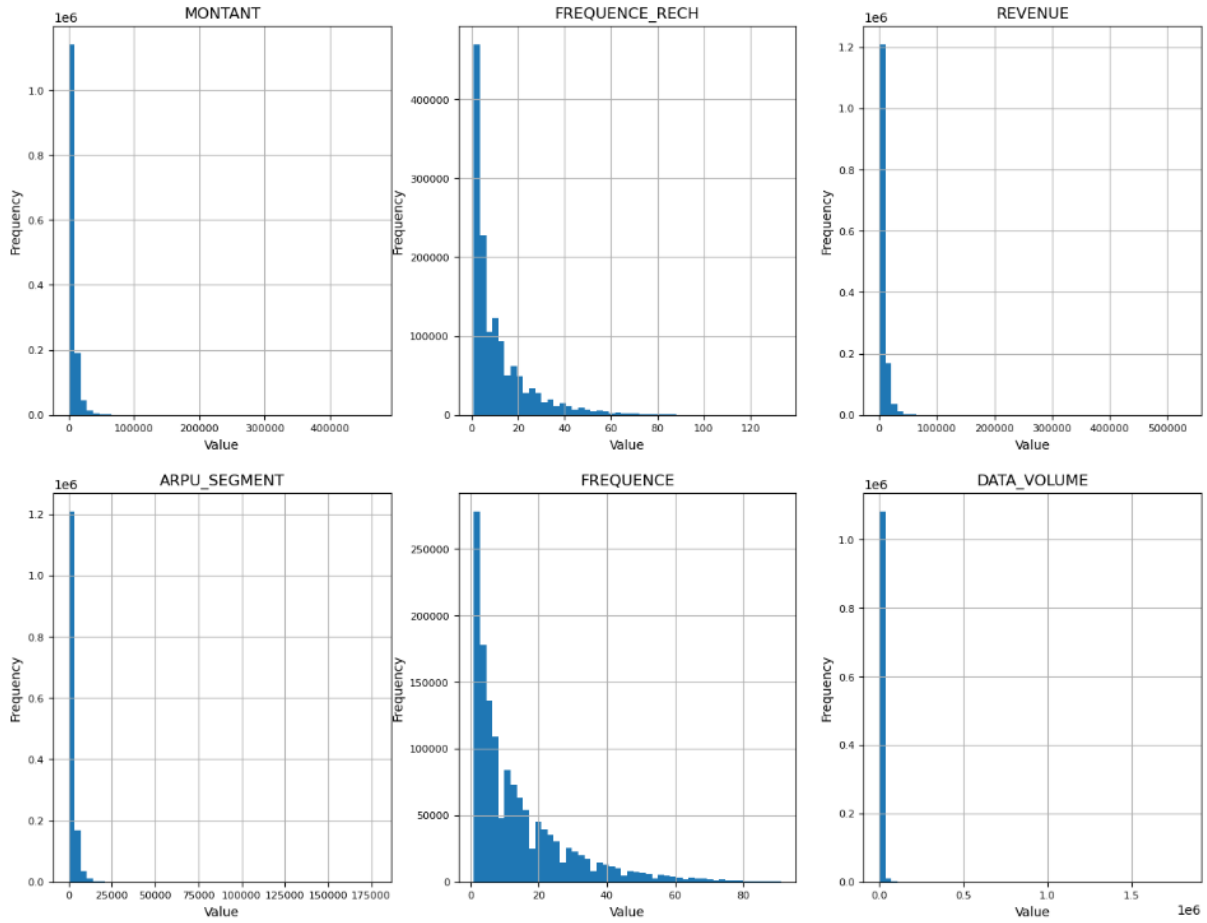
Data Overview and First Captures

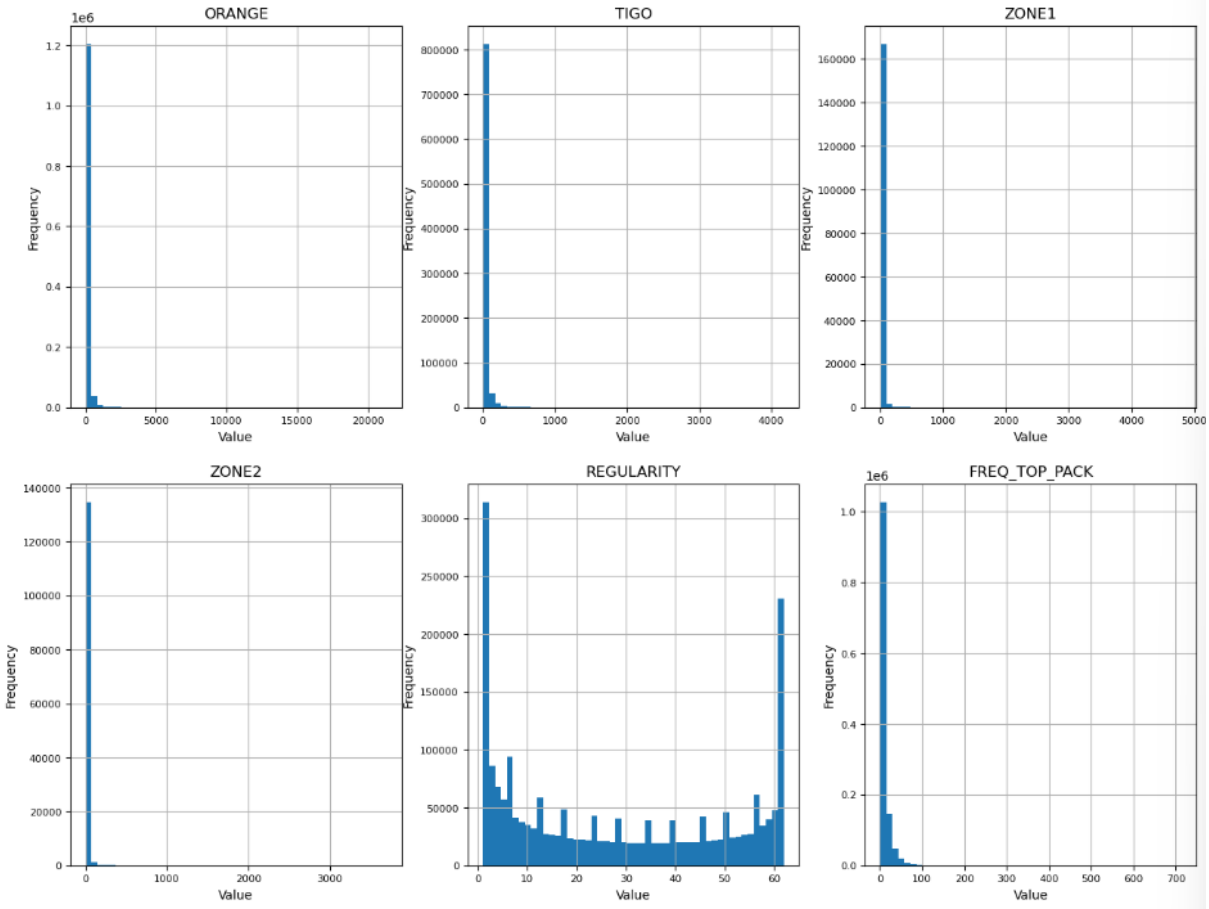
Accurate churn prediction is important for businesses, as customer attrition can impact revenues and overall performance. Artificial Intelligence and Machine Learning techniques are used to achieve this. Our dataset was sourced from the Zindi platform. It includes 19 features and a total of 2154048 records, with the Churn feature as our target column. The dataset has several characteristics that require attention in the analysis. Firstly, there are missing values, which poses a challenge to data completeness. Additionally, the dataset is imbalanced, with certain classes or categories underrepresented compared to others. Moreover, the scale difference among columns introduces complexity that needs addressing during preprocessing. Furthermore, the presence of outliers in certain features highlights the importance of robust data treatment methods. However, the absence of duplicates in the training dataset assures a clean and reliable analytical process. Therefore, a comprehensive data preprocessing strategy is necessary for accurate and insightful model development.

Exploratory Data Analysis (EDA)

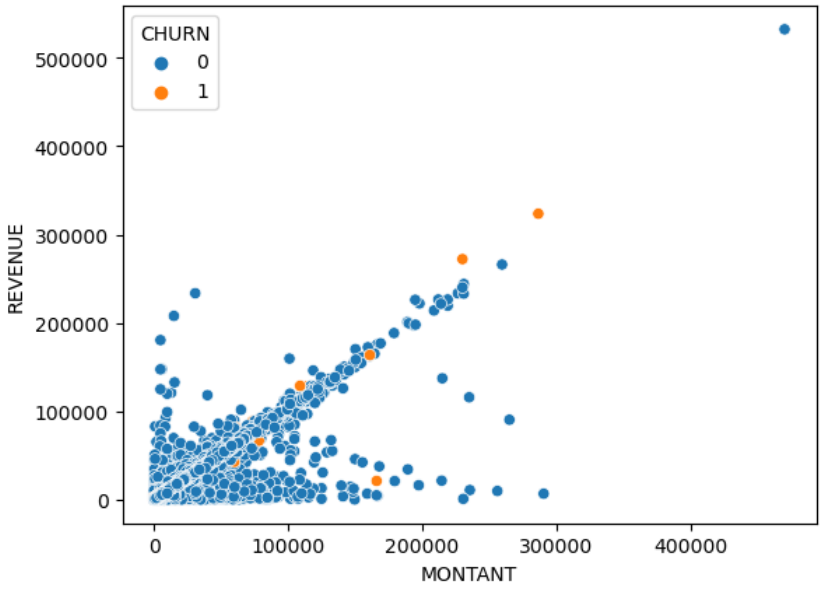
The histograms and box plots provide insights into the dataset patterns. Some variables, such as MONTANT, REVENUE, DATA_VOLUME, ARPU_SEGMENT, ORANGE, TIGO, ZONE1, ZONE2, and FREQ_TOP_PACK, have a lower bound of 0 with a significant frequency,

suggesting that a considerable number of observations fall within this range. This could indicate a common minimum value or threshold for these variables. Conversely, FREQUENCE_RECH and FREQUENCE have right-skewed distributions, indicating the presence of outliers that are higher than the typical range. These outliers may represent exceptional cases or anomalies that require further examination. Furthermore, the variable REGULARITY stands out as a crucial feature that differentiates between retained and churned target user groups. The boxplots are shown below:

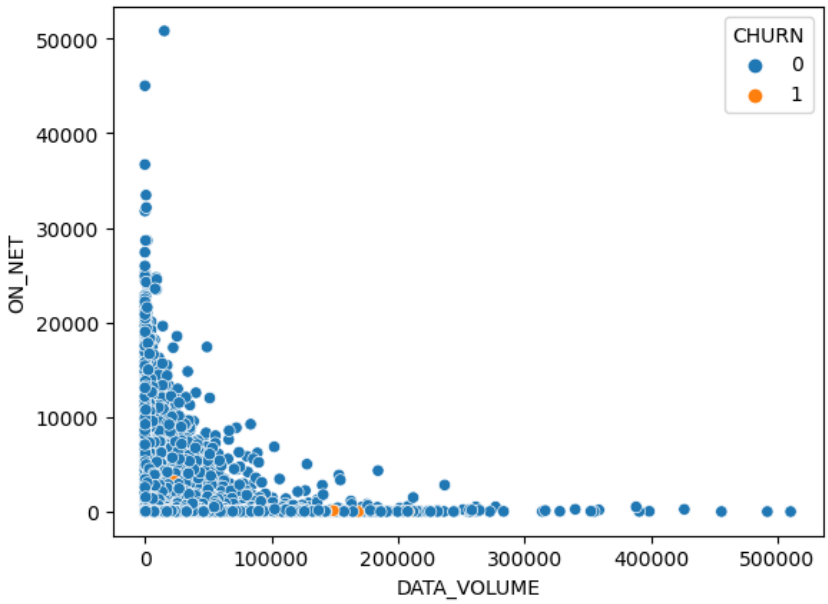




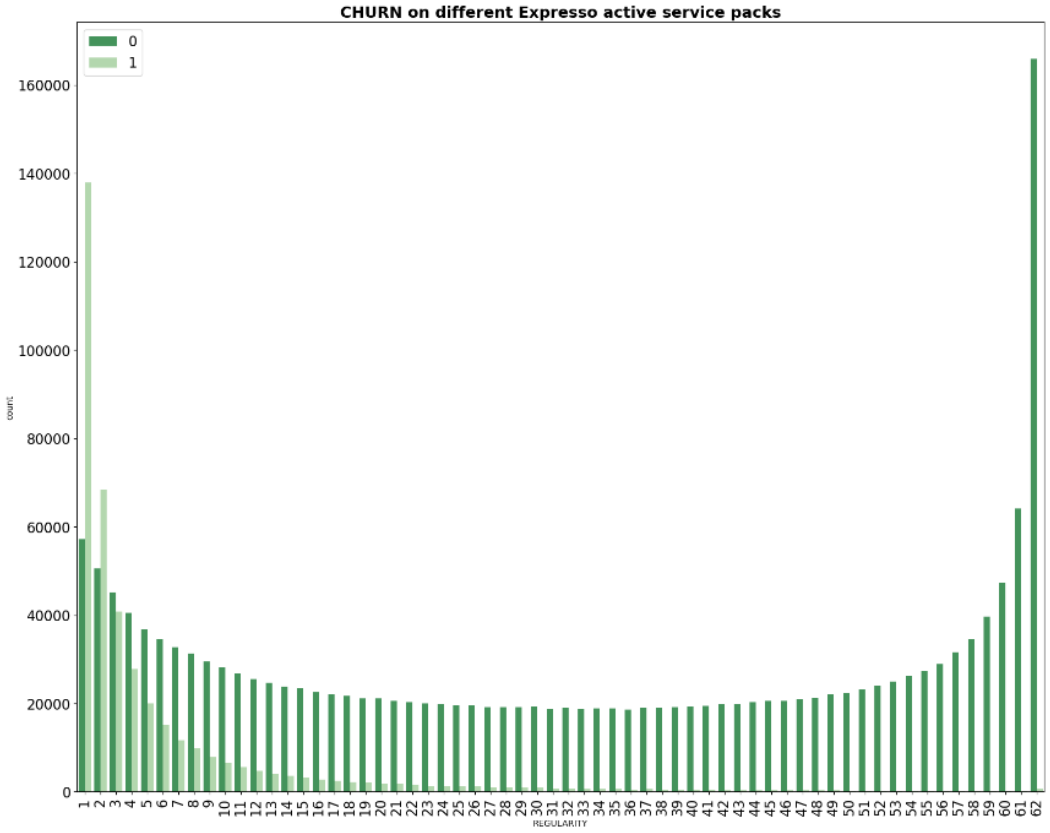
We also categorized our data based on these features, user information, behaviour, payment patterns, and activity indicators. About the user's payment patterns, we found that individuals who make higher payments to Expresso generally have higher monthly incomes, highlighting the importance of considering features such as MONTANT and REVENUE. These features play a crucial role in determining users' willingness to spend on Expresso's products or services relative to their income levels. The scatterplot graph below clearly shows the relationship between REVENUE and MONTANT features:



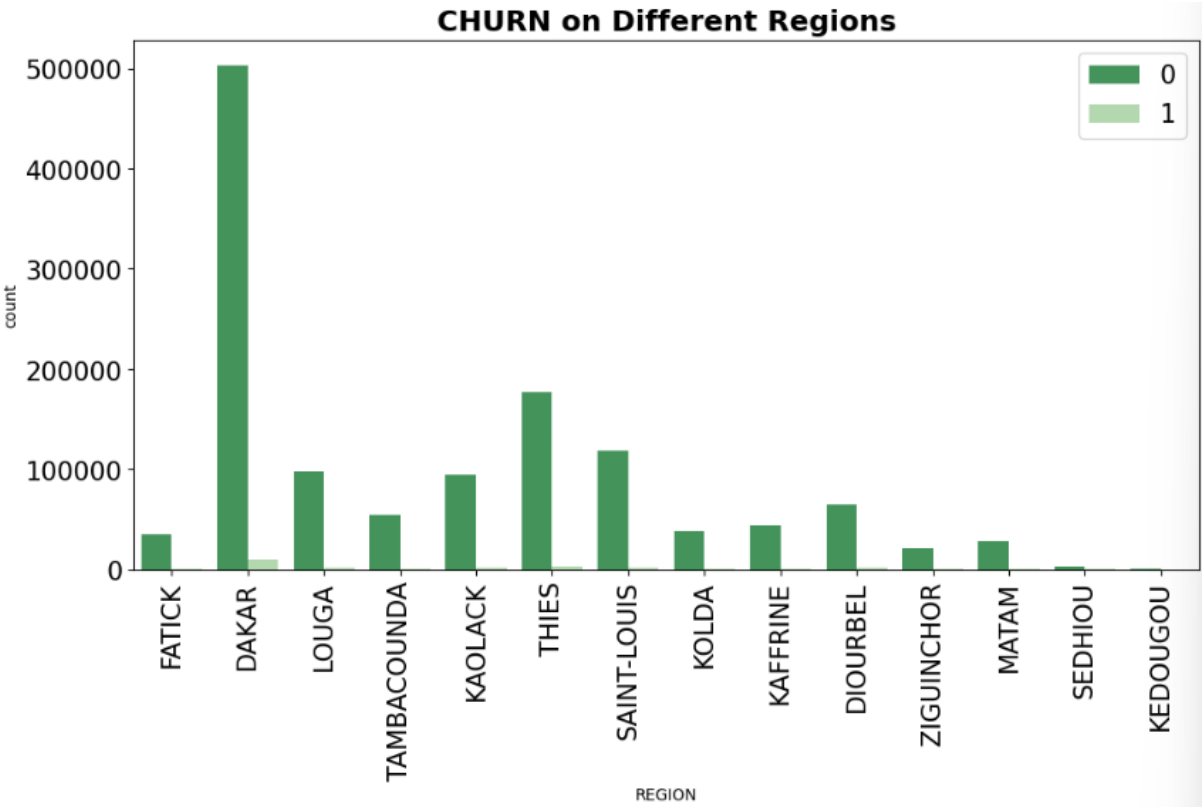
Regarding user behaviour, we found that the usage patterns of the Expresso service are not immediately clear, especially concerning cross-network calls. However, two critical features emerged as significant contributors to our analysis: DATA_VOLUME, which indicates the amount of data used, and ON_NET, the count of inter-Expresso calls made by users. These features suggest that data usage and interaction within the Expresso network are essential in understanding user behaviour.

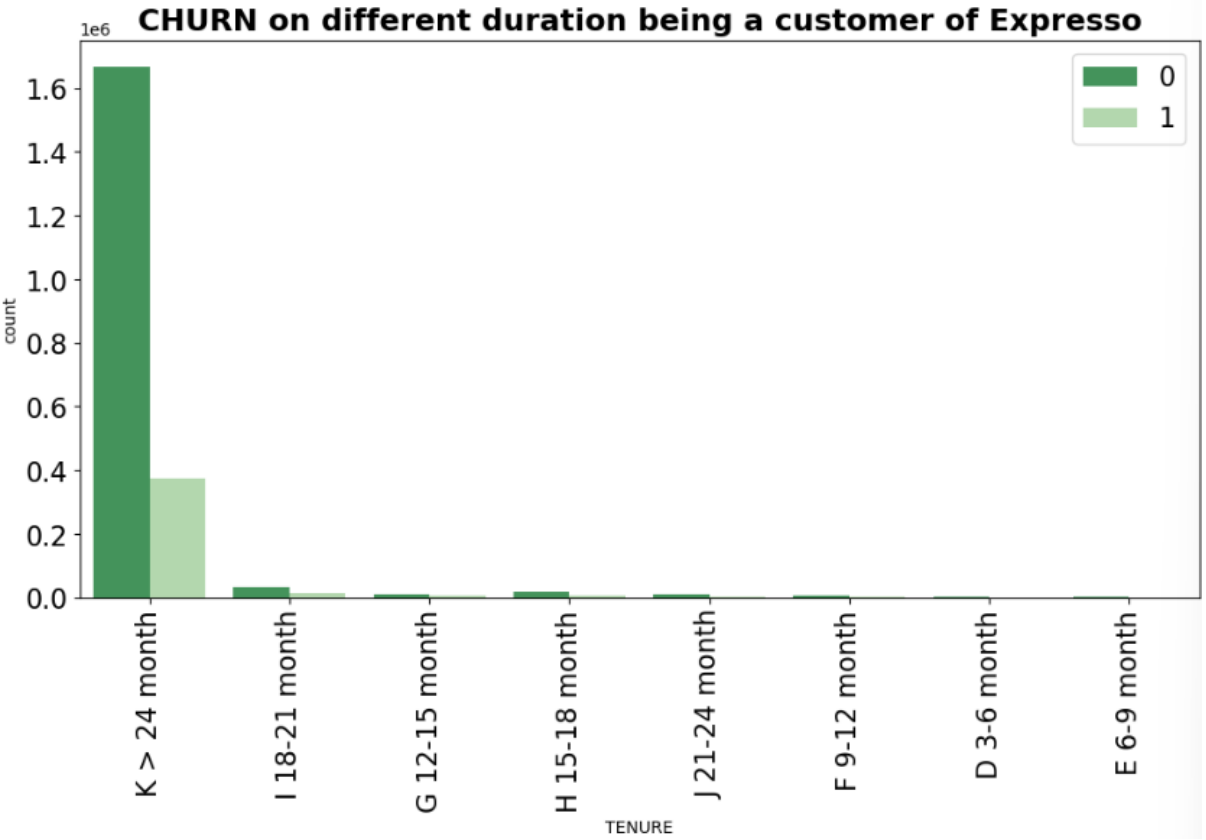


When it comes to user activity indicators, REGULARITY stands out as a central feature that showcases distinctive trends for churned and retained user groups. Users who exhibit prolonged activity over multiple months tend to be retained, while churned users show the opposite trend. Additionally, specific top packs, such as "ALL_NET 500F = 2000F 5d" and "DATA: 490F = 1GB 7d," align with higher churn rates. Therefore, REGULARITY and TOP_PACK become key features in this category, providing valuable insights into user retention and attrition. Overall, these distinct groups based on different features provide insight into the complex interplay between payment behaviours, service usage, and activity indicators within the Expresso user base.



Regarding user's information group of features, there are not many significant insights for specific regions customers would churn, but DAKAR is the region having the highest churn rate among other given regions. Apart from that, users are also likely to churn after being an Expresso member using its products and services about 24 months onwards. There would be features that users are not satisfied after that amount of time and Expresso need to take it into account for other new upcoming users.



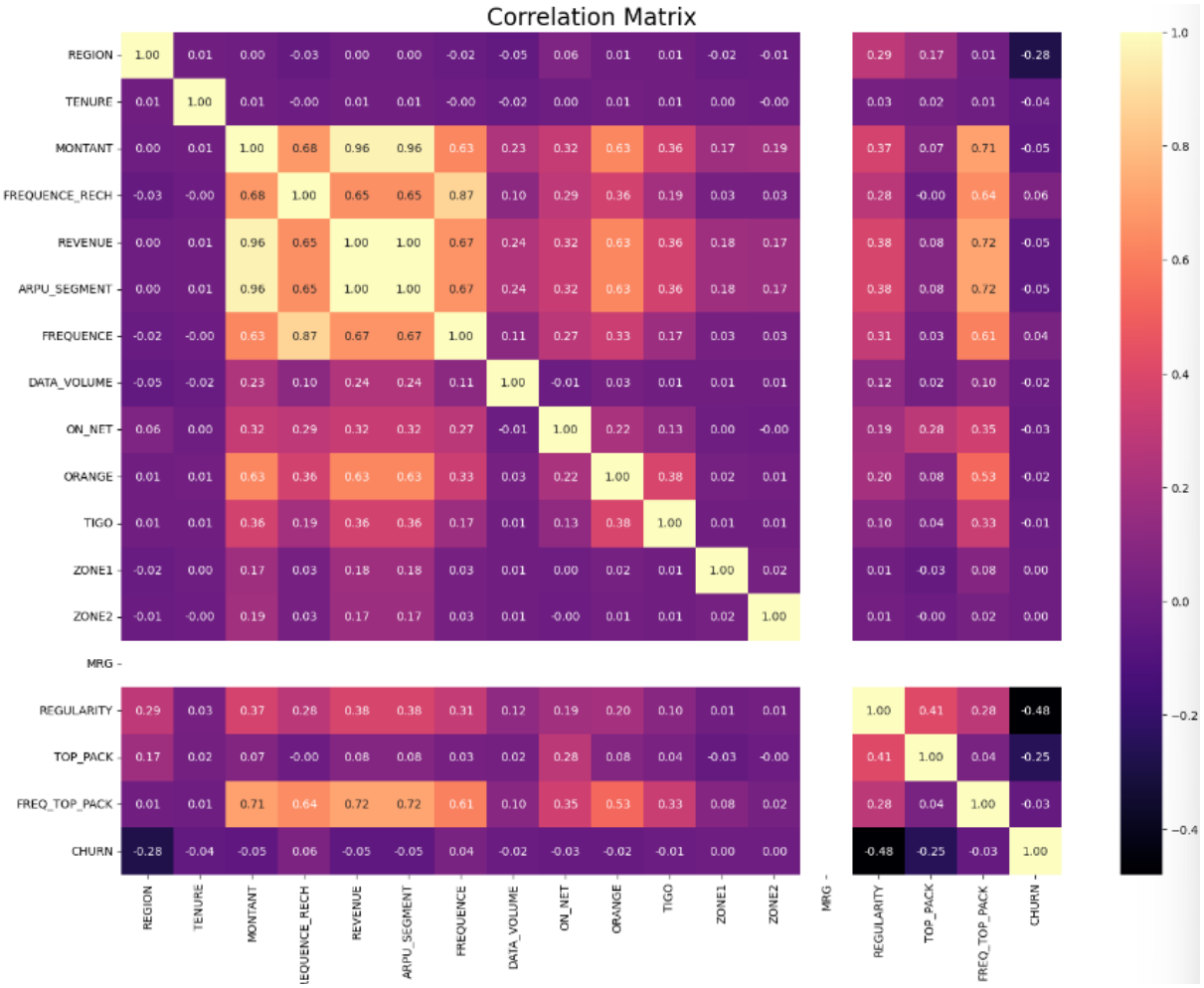


Train, Validation and Test splits and Data Pre-processing

Since our dataset is part of a competition there are no test labels attached hence we decided to split the dataset into three parts: Train, Validation and Test sets. Data processing was required to transform both numerical and categorical columns. Transformation on numerical columns such as filling missing values, removing duplicates, logarithmic transformation to correct skewness and converting categorical variables using label encoder, and scaling are all defined into a single function and we pass the split data sets through these pipelines individually to ensure no bias arises. The whole data pre-processing is packed in the full pipeline for further analysis when new data is fed in the project for predicting the churn rate. In term of solving the imbalanced issue in our dataset, using SMOTE technique is an appropriate technique where the minority class is imputed with frequent values using spatial nearest neighbour arrangement.

Correlation Analysis with Target Churn Column

Using the correlation matrix we cross-confirm with our EDA that the top related variables are REGULARITY, TOP_PACK and REGION (negatively correlated with the Churn target column) and decide to use all the variables to not compromise on the additional insights.



Building Different Models and Best Model Selection

The dataset is trained and tested using different models, namely Logistic Regression (Our Baseline Model for Model Selection), Random Forest, Gradient Boosting Classifier, and Extreme Gradient Boosting (XGB) classifier. Based on a range of performance evaluation metrics including, Precision, Recall, Accuracy Score, F1 Score, and ROC AUC Score, Random Forest out of all the models has provided the best predictions and was validated using the

Accuracy and ROC AUC Scores of 0.8077 and 0.8079. The first runner-up is our baseline model, Logistic Regression. Therefore, we decided to move forwards with Random Forest Classifier and Logistic Regression for further improvement. Here are the table of results from all models:

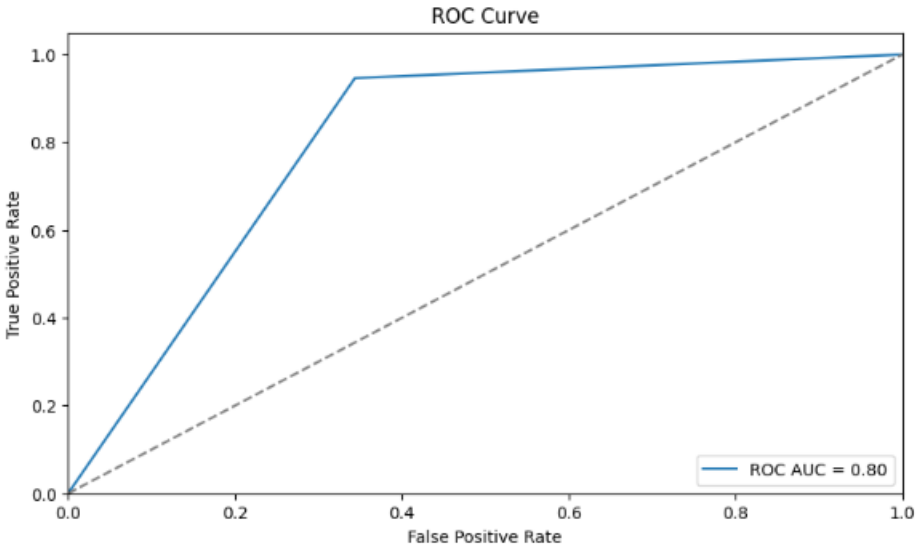
Algorithm	Accuracy Score	ROC AUC Score
Logistic Regression	0.8040683916410216	0.8043494182609174
XGB	0.5376374633576532	0.5384540409220981
Random Forest	0.8067109258074807	0.8069671130709775
AdaBoost	0.5	0.4991151875088283
GBDT	0.7169822101108277	0.7174689210247123

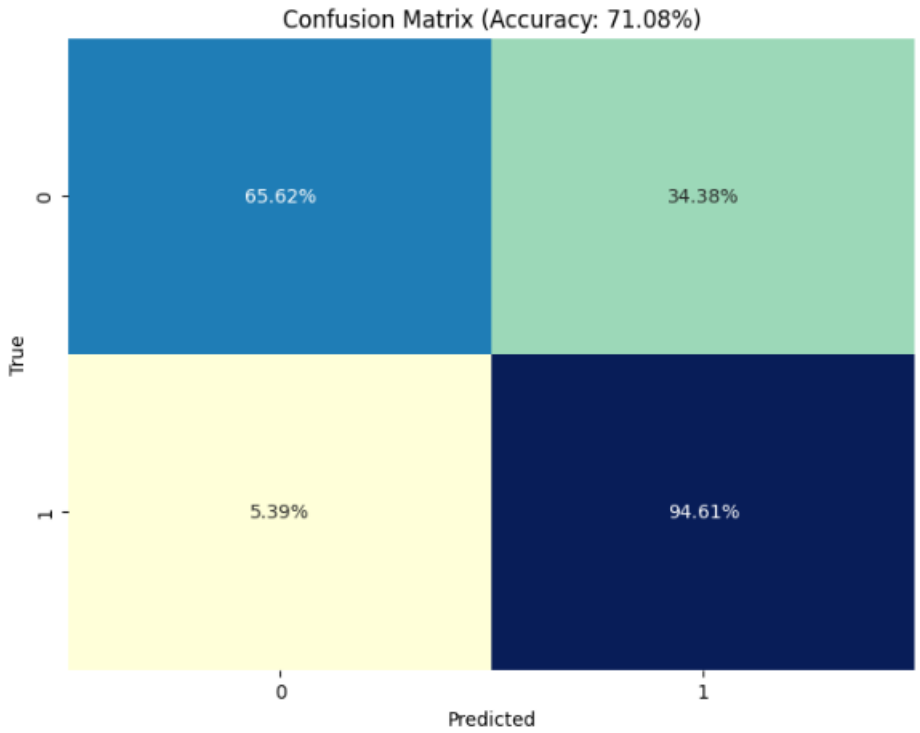
Random Forest Model's Optimization

The next step is that we randomly select the parameters' value for Random Forest Classifier and see how improved the model would be. Those parameters are max_depth, n_estimators, in_samples_split which are significantly important for Random Forest to maximize its performance. As a result of that set of chosen parameters, we observed a considerable improvement in those evaluation metrics, specifically the Accuracy and ROC AUC Scores of 0.8445 and 0.8446. Therefore, it is clearly that hyperparameter tuning should be considered to further improve the model and results.

Results

Based on our extensive model training and evaluation, we discovered that the Random Forest model performed exceptionally well, achieving a remarkable ROC AUC score of 0.8446. By wanting to check the goodness of the Random Forest Model to confirm that the model is not overfitting and biased, we conducted the cross-validation with the number of K fold of 3. The results on Average Accuracy and ROC Scores on the entire training data were 08586 ad 0.9243 respectively. The final step is to apply the best model – Random Forest Classifier on the test data (x_test and y_test) we have separated previously. Here are the ROC Curve Graph and Confusion Matrix:





Our thorough analysis also revealed that the key factors that significantly influence customer churn are Regularity, Region, Top-Pack and Montant. By taking the necessary steps to address these factors, companies can retain customers and prevent revenue loss. With this valuable information, companies can now focus on developing targeted strategies to address the identified factors, which, in turn, can help them retain their customers and improve their bottom line. Regularity, for instance, can be addressed by offering a personalized rewards program to encourage customers to stick with the company. Region, on the other hand, can be addressed by tailoring marketing and promotional activities to reflect the unique needs and preferences of customers in specific regions. Finally, Montant can be addressed by offering flexible pricing plans that cater to the different budgets of customers.

Conclusion and Future Works

The purpose of this project is to help the company identify which customers are at risk of churning, so that the company can take appropriate actions to retain them. We have used ML model to flag them, which would be very helpful for the marketing team to take necessary operation actions.

After our exploratory data analysis, we got basic insights about our data. Then we preprocessed the data, and tried many supervised learning approaches, such as Logistic Regression, AdaBoost, and Gradient Boosting Classifier. We evaluated the performance of each model using the ROC AUC metric for the Random Forest Classifier. Then, we used the best model to make predictions.

After training and evaluating our models, we found that the Random Forest model had the best performance with a ROC AUC score of 0.844 with a set of selected parameters. We also found that the most important features for predicting customer churn were REGULARITY, REGION, MONTANT, and TOP_PACK, which can help the company take appropriate actions to retain customers and avoid loss of revenue.

There are ways to improve our predictive analysis methods in the future. One such method is hyperparameter tuning, which has shown improvements in our results and has the potential to enhance the accuracy of our predictive models. Another way to refine our models and gain insights is by exploring feature importance and implementing a feature selection process. It's also intriguing to abstain from using the SMOTE technique to fully assess its impact on our results and potentially uncover new strategies to improve our

predictions. By working together on these efforts, we hope to take our predictive analysis to new levels of effectiveness and dependability.

References

https://www.kaggle.com/datasets/hamzaghanmi/expresso-churn-prediction-challenge

https://zindi.africa/competitions/expresso-churn-prediction