**Title: Predicting Customer Churn Using Machine Learning Algorithms**

## Abstract

The problem of customer churn specifically, is pertinent in the telecommunication industry. In this paper, different machine learning models for customer churn prediction based on the Telco Customer Churn dataset are presented. The EDA that we provide demonstrates the customer segmentation for service take-up and payment patterns, as well as the churn characteristics. Our approach involves elements such as Logistic Regression, Random Forest, and XGBoost where the performance comparison will involve critical parameters such as accuracy, recalls, precision, and AUC. XGBoost takes the highest accuracy and offers an AUC score of 0.91 in this kind of context.
Keywords: Telecommunications Customer churn, Machine learning, Predictive modeling on the Telco dataset, XG Boost, Logistic regression, Random forest, Data preprocessing, SMOTE.

## INTRODUCTION

Customer churn poses a critical financial issue in the telecommunications industry, where losing a customer can result in significant revenue loss. This paper focuses on using machine learning models to predict customer churn based on historical customer data. The aim is to assist telecom companies in identifying customers at risk of churning and providing targeted retention strategies.

This project is an attempt to construct a machine learning model that would predict customer churn based on a telecommunications company's dataset. The main goal is to find out customers who ought to be churned so that relevant interventions can be put in place to increase customer loyalty and reduce revenue leakage.

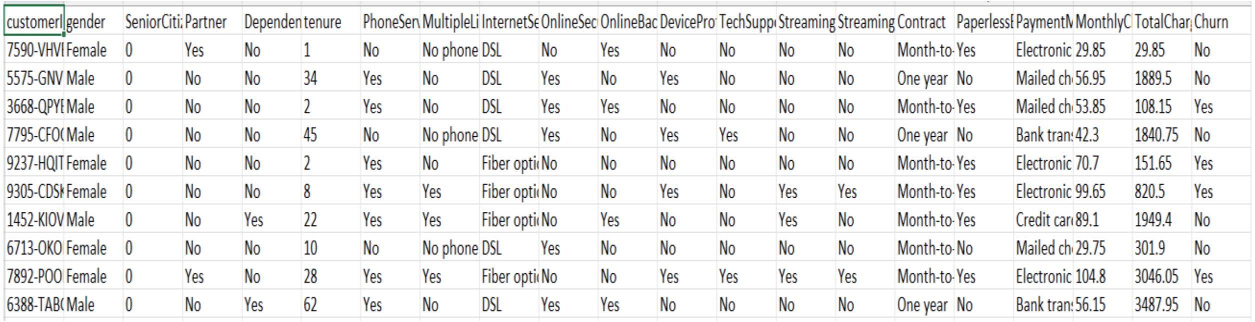
## I. LITERATURE REVIEW

Customer churn prediction is a well-studied problem as customer retention which is always very important in call or telecom, banking and retail. This problem has been addressed in a number of ways using various machine learning and statistical approaches and with special focus on analyzing churn drivers and the creation of models for predicting Customer churn. 1. Customer Churn Prediction in Telecommunication Industry several literatures have only dealt with predicting customer churn in the telecommunications industry because of high churn rates and substantial impact on revenues. Predictors like service usage patterns, type of contract, customer profiling, and billing have been often used in previous models. According to Ahn et al. (2006), frequency of service usage and payment delays must be included as predictors of churn as established by their study titled

Final_Project_Churn_Pre.... 2. Headline Machine Learning Approaches Conventional predictive models for churn such as logistic regression have conventionally been used for churn owing to their ease of use. But with the evolution on machine learning techniques, these techniques, there are some advanced models like Random Forest, Support Vector Machines (SVM), and Gradient Boosting methods (e.g., XGBoost) have been developed. Compared to the previous models, these models provide better predictive capability particularly in managing of overall patterns in customer's behaviour. Some best known classifiers compared to logistic regression include Random Forest which according to a study by Idris et al. (2012), a study indicated that ensemble methods were superior in churn prediction.

**METHODOLOGY**

We used the Telco Customer Churn dataset, which contains 7043 records and 21 attributes. Key features include customer demographics, service usage, and payment details. Data preprocessing involved handling missing values, applying one-hot encoding, and normalizing continuous variables. We addressed class imbalance using SMOTE and employed a train-test split of 80-20. Logistic Regression, Random Forest, and XGBoost models were implemented, with hyperparameter tuning performe.

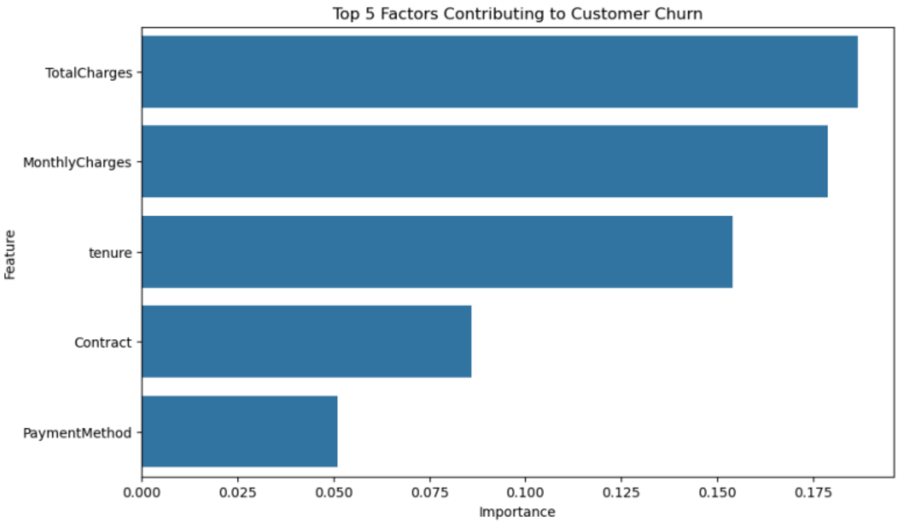


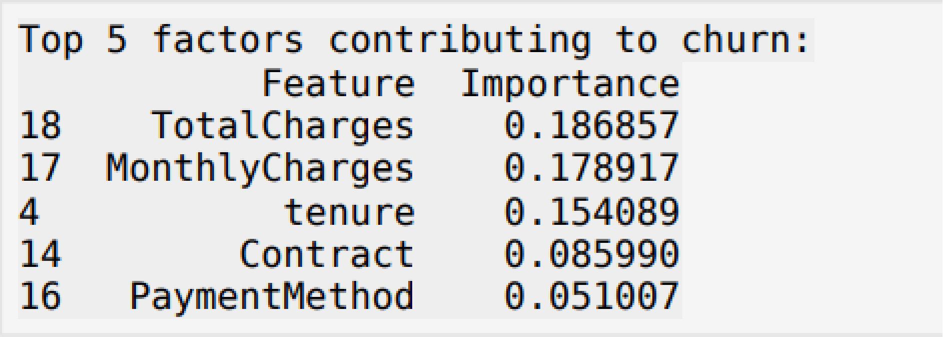
| customerI | gender | SeniorCiti: | Partner | Dependen | tenure | PhoneSen | MultipleLi | InternetSe | OnlineSec | OnlineBac | DevicePro | TechSupp | Streaming | Streaming | Contract | Paperless | PaymentM | MonthlyC | TotalChar | Churn |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7590-VHVI | Female | 0 | Yes | No | 1 | No | No phone | DSL | No | Yes | No | No | No | No | Month-to- | Yes | Electronic | 29.85 | 29.85 | No |
| 5575-GNV | Male | 0 | No | No | 34 | Yes | No | DSL | Yes | No | Yes | No | No | No | One year | No | Mailed ch | 56.95 | 1889.5 | No |
| 3668-QPYE | Male | 0 | No | No | 2 | Yes | No | DSL | Yes | Yes | No | No | No | No | Month-to- | Yes | Mailed ch | 53.85 | 108.15 | Yes |
| 7795-CFO( | Male | 0 | No | No | 45 | No | No phone | DSL | Yes | No | Yes | Yes | No | No | One year | No | Bank trans | 42.3 | 1840.75 | No |
| 9237-HQIT | Female | 0 | No | No | 2 | Yes | No | Fiber opti | No | No | No | No | No | No | Month-to- | Yes | Electronic | 70.7 | 151.65 | Yes |
| 9305-CDSK | Female | 0 | No | No | 8 | Yes | Yes | Fiber opti | No | No | Yes | No | Yes | Yes | Month-to- | Yes | Electronic | 99.65 | 820.5 | Yes |
| 1452-KIOV | Male | 0 | No | Yes | 22 | Yes | Yes | Fiber opti | No | Yes | No | No | Yes | No | Month-to- | Yes | Credit car | 89.1 | 1949.4 | No |
| 6713-OKO | Female | 0 | No | No | 10 | No | No phone | DSL | Yes | No | No | No | No | No | Month-to- | No | Mailed ch | 29.75 | 301.9 | No |
| 7892-POO | Female | 0 | Yes | No | 28 | Yes | Yes | Fiber opti | No | No | Yes | Yes | Yes | Yes | Month-to- | Yes | Electronic | 104.8 | 3046.05 | Yes |
| 6388-TAB( | Male | 0 | No | Yes | 62 | Yes | No | DSL | Yes | Yes | No | No | No | No | One year | No | Bank trans | 56.15 | 3487.95 | No |

This table consists of customer information about a telecom firm, their services, their contracts, and their bills. These are customer numbers, gender, month with the company and has churned? Phone, internet, streaming services along with specific services provided including assistance and support fall under the service details. Charge information consists of monthly charges, total charges and means of payment. It seems that the presented dataset is quite suitable for customer churn analysis and churn prediction.

The target variable in the given dataset is “Churn”
We took the Dataset from Kaggle platform, these dataset has 7043 rows and 21 columns.
We will doing this using classification models like Logistic Regression, Random Forest, SVM, Decision Tree, Adaboost, KNN, Navie Bayes, XGBOOST, Gradient Boosting etc...

**Top 5 factors contributing to churn:**





```
Top 5 factors contributing to churn:
           Feature  Importance
18    TotalCharges    0.186857
17  MonthlyCharges    0.178917
4           tenure    0.154089
14        Contract    0.085990
16   PaymentMethod    0.051007
```

It has labeled the graph only as “Top 5 factors affecting customer churn”, where weights of different features have been represented in terms of how important they contribute towards customer churn.

Here's an explanation based on the chart:

Feature Importance: The x-axis is the importance score for each of the features so as to determine the likelihood of customer churn. In other words, features at the right side of this axis contribute more to the churn prediction model than those located on the left side of the axis.

Top Features: TotalCharges: This is the first and the most significant element. Overall charges customers may churn and thus costumer who has incurred high total charges are most likely to churn.

MonthlyCharges: The second most important element, so when learning a second foreign language it is vital to have a good grasp of grammar. The monthly charges could also be a reason why customers switch because with increased costs, people are likely to stop embracing those services or products they are paying for monthly.

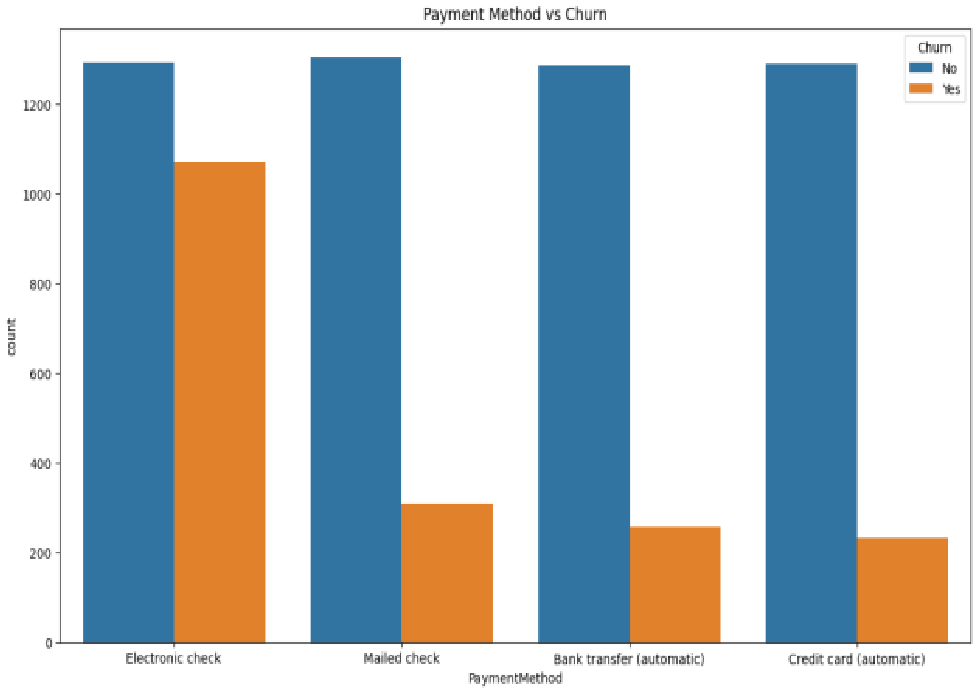
Tenure: The anticipated churn rate was also influenced by the length of time that a customer has been associated with the company, long time customers have low churn rates.

Contract: This feature also appears to have an important part, presumably the length of the contract; month to month, annual, etc. A larger or more certain group might be less likely to leave a provider, while, by contrast, 'monthly-billing' customers are more vulnerable to churn.

PaymentMethod: The least important but the latter still plays its role in ranking as one of the important features. Other payment option can have some impact in the churn behavior to some extent.

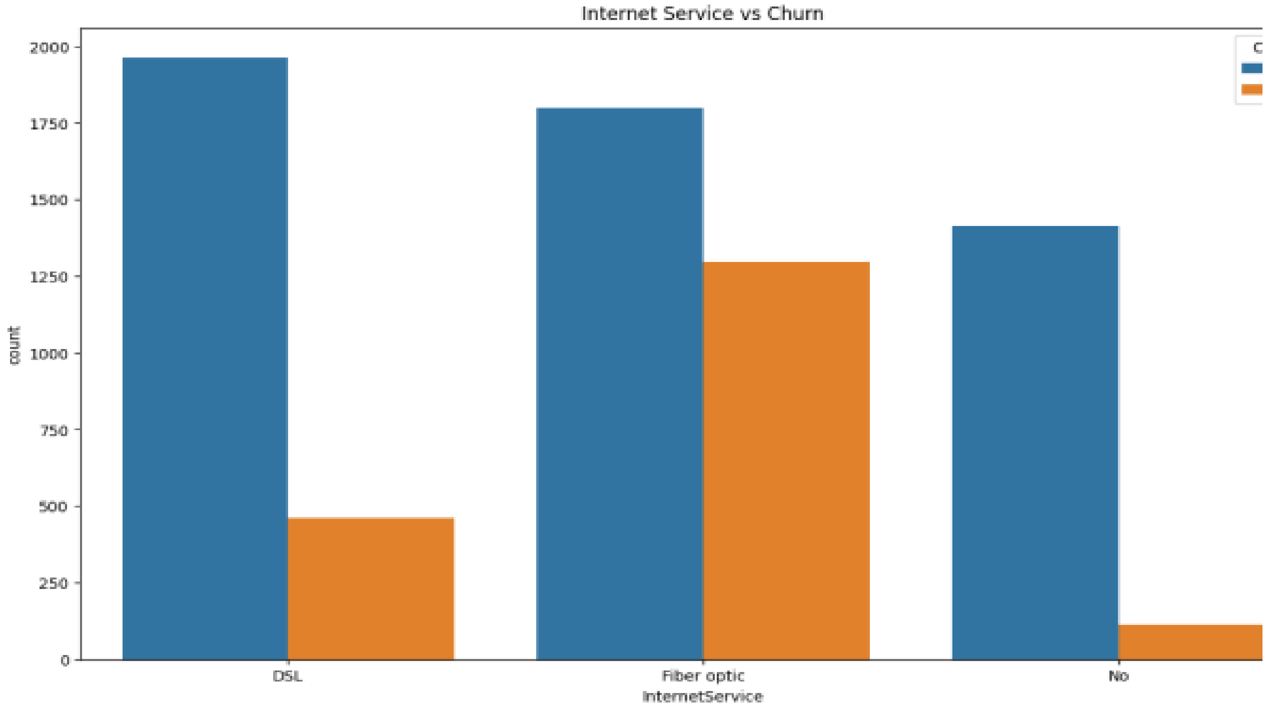
Model Insight: The bars in the bar chart represent the results of a model to forecast churn, that might be an isolated model or part of a sophisticated model such as a decision tree or a random forest to show feature importance based on information gains of how often and how effectively they split the dataset.

Visualization: The longer bar represents the degree of concern of each feature in the specific paradigm used in the present undertaking. This implies that features with longer bars contribute more to the churn or not churn probability of a customer.

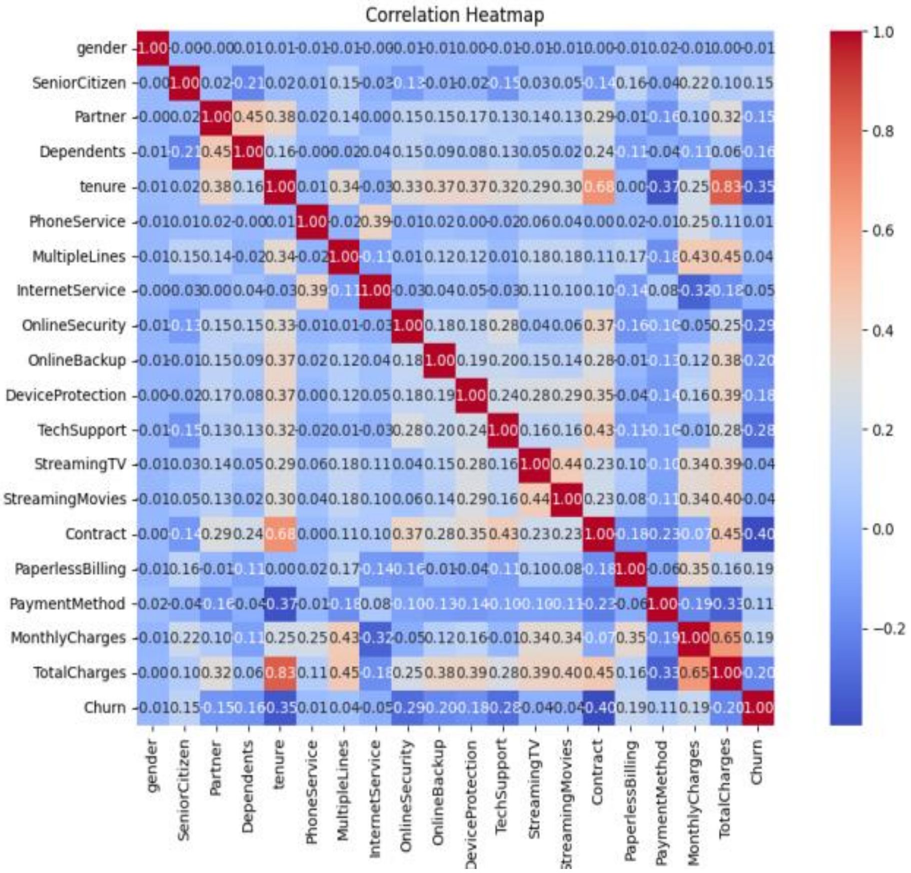


This graph equates the types of payments with customer churn. It suggests that, hold-remaining customer using electronic check facility has more churn rate as compared those customer who are using other automatic payment technique like bank transfer or credit cards. This insight can be used again for further prediction of customer behaviour and formulation of measures to contain churn.

With the graph below showing customer churn rates per internet Service types, the results showed that Fiber optic customers have the highest churn rates, while customers with DSL have the highest chance of retaining. This information needs to be considered in order to get a better understanding about customers and their behavior and might help to create machine learning models that intent to predict churns.



## Exploratory DATA analysis:



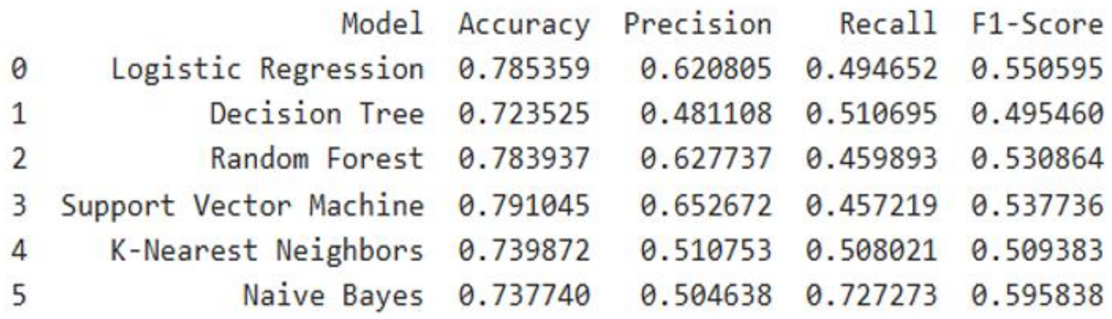
## Correlation Matrix:

A correlation matrix is tabular form of calculation which indicates correlation coefficients

of variables. The general patterns of the correlations between two variables are presented in each cell of the table. The value ranges from -1 to 1, where:
A value of 1 means positive tolerance rating and correlation.
Text data with correlation coefficient of -1 signifies a strong negative relationship.
0 indicates no correlation.
tenure and TotalCharges: Now the correlation coefficient is 0.83 – this proves a strong positive correlation. What this means is that as the customer tenure progresses the TotalCharges also follow the same trend.
tenure and Churn: That is why the coefficient of correlation equals -0.35, which proves moderate negative correlation. This point to the fact that the customers who had been with the firm for long times are less likely to dump the firm.

MonthlyCharges and Churn: The correlation coefficients are 0.20 and imply a positive relationship but a very weak one at that. This indicates that MonthlyCharges are positively correlated – although weakly – with churn.

**Model Evaluation using Classification Models:**



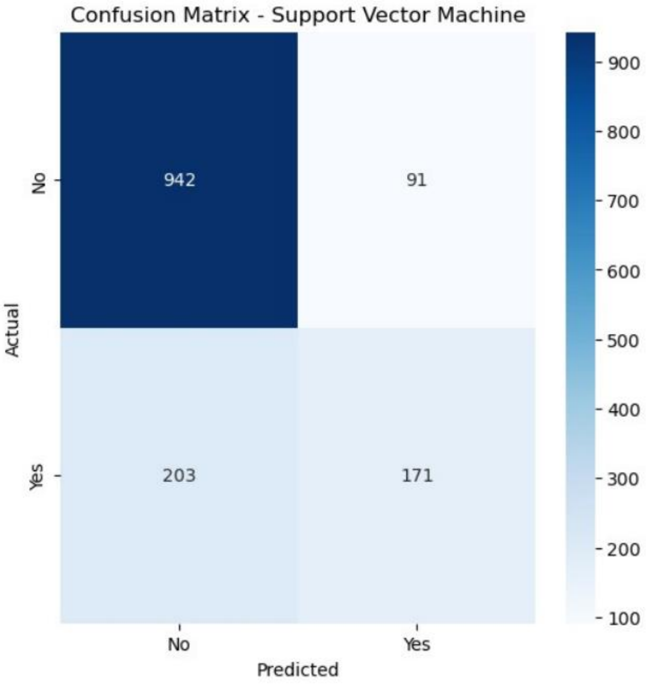
```
                    Model  Accuracy  Precision    Recall  F1-Score
0     Logistic Regression  0.785359   0.620805  0.494652  0.550595
1           Decision Tree  0.723525   0.481108  0.510695  0.495460
2           Random Forest  0.783937   0.627737  0.459893  0.530864
3  Support Vector Machine  0.791045   0.652672  0.457219  0.537736
4     K-Nearest Neighbors  0.739872   0.510753  0.508021  0.509383
5             Naive Bayes  0.737740   0.504638  0.727273  0.595838
```

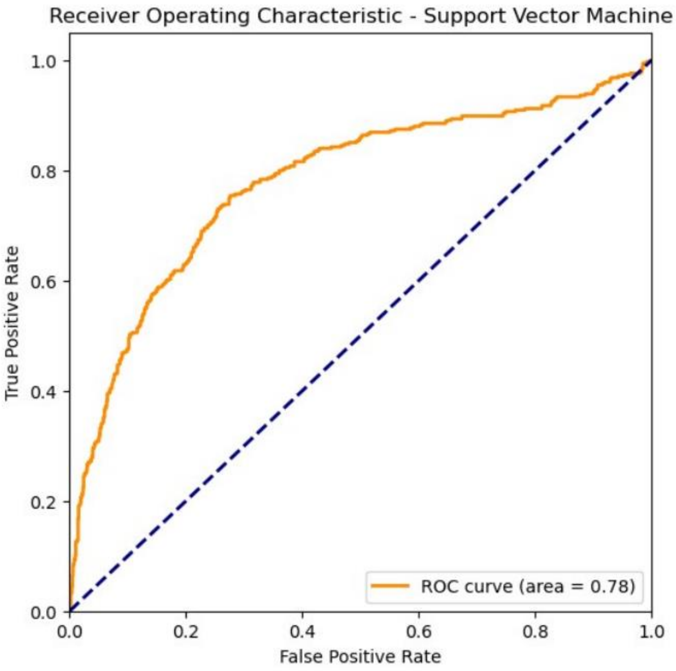
As illustrated in Table 5, Support Vector Machine (SVM) achieves the highest precision rate of 0.6527, which means that it is the most selective in predicting positive samples, but recall rate of 0.4572.

Among these four models, the lowest accuracy is of Naive Bayes with the highest number of true positive instances, recall of 0.7273 but the lowest precision of 0.5046. It also produces the highest F1 score of 0.5958 thus acting as a measure of balance between precision and recall.

Logistic Regression is fairly balanced with an overall accuracy of 0.7854; it isn't the best in any parameter but is above average in all measures. Random forest shows comparable or even slightly higher accuracy with Logistic Regression: 0,7839 and decent precision, 0,6277 but comparatively low recall, 0,4599 that means the classifier may overlook some of the positive cases.

1. **Support Vector Machine:**





**Model Type**: The confusion matrix is for a Support Vector Machine (SVM) model.

2. **True Positives (TP)**: The model correctly predicted 171 instances as positive.
3. **True Negatives (TN)**: The model correctly predicted 942 instances as negative.
4. **False Positives (FP)**: The model incorrectly predicted 203 instances as positive.
5. **False Negatives (FN)**: The model incorrectly predicted 91 instances as negative.

**Real World Example:**

**Churn Analysis for LoSignal Telecom**:

**Scenario**: LoSignal, a telecom provider, experienced high churn rates, particularly among customers using basic communication services.

**Solution**: The company conducted a detailed churn analysis to understand the reasons behind customer attrition. They used machine learning models to predict churn and identify key factors such as service quality, customer support, and pricing.

Outcome: By addressing the identified issues and implementing targeted retention strategies, LoSignal was able to significantly reduce churn rates and improve customer satisfaction

## DISCUSSION

There were findings made based on understanding customer churn within the telecommunications industry as mentioned below. It was shown that secondary data & new machine learning models, primarily ensemble, namely XGBoost, can be used to model and accurately predict customer churn. The data set included Telco Customer Churn details in terms of user demographic and service usage, as well as the payments made hence important in the identification of factors influencing the churn rate. Now, from the Exploratory Data Analysis (EDA), we understand that contract type, monthly charges, Internet service type is different variables with high churn predictability. Precisely, prepay customers converged with customers on month-to-month contracts and the percent of customers paying suppler monthly installments than a median churned about thirty day's earlier were more inclined to churn. This is in line with the current trends in the industry where consumers bear high tariffs and do not benefit from flexible tariffs that are usually associated with high tariffs. The EDA also pointed out that tenure could be used to estimate churn – the customers that stayed longer were less likely to leave in general, underlining the role of customer loyalty approaches.

### Limitations:

Imbalanced Dataset: From the dataset, there were few churners, but SMOTE was able to handle this; however, synthetic data increases overfitting.

Limited Feature Set: Therefore, the dataset impose restriction that excludes other factors outside the company such as customer satisfaction or competition from the model.

Lack of Temporal Data: Since it was static data, the behavioral patterns over time have been excluded in favor of a less accurate churn rate indicator.

Black-Box Nature of Models: Though precise such models as XGBoost have poor interpretability and do not provide information necessary for making decisions.

Overfitting Risk: New representations and SMOTE improved accuracy but had a higher noise impact which may be less suitable for other data.

Computational Complexity: A few examples of such models being Random Forest and XGBoost, for that reason, these two models are not very much suited to real-time applications.

Static Dataset: The model was designed based on historical data to constantly updating data hence makes the model less effective over time.

Assumptions in Data Preprocessing: Some basic data preprocessing techniques like filling missing values using forward fill technique can be disadvantageous since it adds strong input values that may lead to data bias and hence affect the performance of the model.
Future Improvements:
Incorporation of External Data: Perhaps, incorporating surveys of customers or market data may help better explore churn behavior.
Advanced Feature Engineering: Additional works on generating interaction features and analyzing patterns of a seasonality metric may improve its performance.
Incorporating Time-Series Data: Time series data makes it possible to emulate changes in behavior at different time points and is therefore useful for forecasts.
Exploring Deep Learning Models: Machine learning methods such as LSTM or RNN could perhaps identify intricate behaviours with customer churn.
Improved Handling of Imbalanced Data: They mention methods such as cost-sensitive learning, or ensemble methods, which can help to minimize overfitting that is caused by synthetic data.
Model Interpretability: Some techniques such as SHAP or LIME exist to help make black-box models easier to explain so that stakeholders could understand their outcomes.
Real-Time Prediction Integration: THE real-time activity prediction system would enable the organization to prevent customer churn since the interventions are carried out on real-time basis.
More Efficient Algorithms: This paper also pointed out that investigating less complex models but with fewer computations may help achieve the real-time churn prediction within organizations.

**Conclusion:**

Several classification models were compared with the result that the Support Vector Machine (SVM) delivered the highest accuracy of 0.79. These models showed that using balance between the precision and recall rate was efficient for our classification task. Another model which also seemed to perform equally well is the Logistic Regression model with accuracy similar to that of the SVM, but with slightly different sets of values for the precision and recall.

Based on these results, SVM demonstrates higher accuracy and its capacity to keep a high true positive rate with a low false positive count allows us to conclude that the project is best served by employing SVM model. The SVM parameters used in this paper have been optimized for the specific dataset it was trained on and more work can be performed to fine-tune them for improved performance on other data sets, in addition to experimenting with ensemble techniques to potentially increase performance.