



SYRIATEL
COMPANY



TABLE OF CONTENT

- INTRODUCTION
- BUSINESS UNDERSTANDING
- OBJECTIVE
- EXPLORATORY DATA ANALYSIS
- PREPARATION FOR MACHINE LEARNING
- CONCLUSION AND RECOMMENDATION

INTRODUCTION

This research employs machine learning algorithms to create a model that can accurately predict which customers will churn based on the information in the dataset. The dataset includes 20 predictor variables, the majority of which relate to client usage habits. The goal variable is 'churn'. Because the goal variable is categorical, classification methods are utilized to develop the prediction model. Recall is used to assess the model's performance.

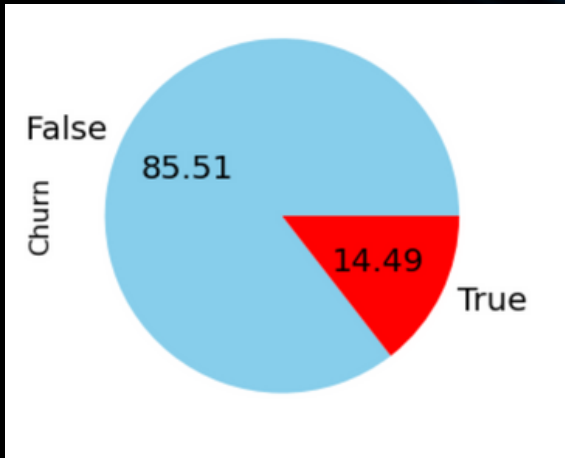
BUSINESS UNDERSTANDING

To strengthen their revenue base, telecommunications businesses must recruit new consumers while increasing client retention. Syriatel is a mobile telecommunications and internet services company headquartered in Damascus, Syria. It has been demonstrated that maintaining long-term client connections is more effective than attempting to recruit new customers. Churn prediction has thus become a crucial component of the company's strategy. The goal of this project is to create a model that accurately predicts which customers are most likely to churn, as well as to determine the characteristics that are significant for predicting customer churn. Syriatel can thus act to prevent churn from occurring.

OBJECTIVE

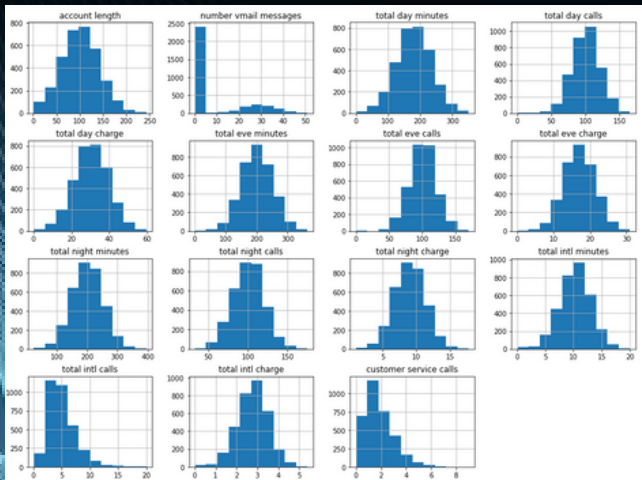
- To build a machine learning model that can accurately predict customers who will churn based on the information available in the dataset.
- To identify the features that are important for predicting customer churn.

EXPLORATORY DATA ANALYSIS




The target variable for this classification project is "churn".

There is a class imbalance problem since the target class has an uneven distribution of observations. 85.51% of the data belongs to the False class while 14.49% belongs to the true class.



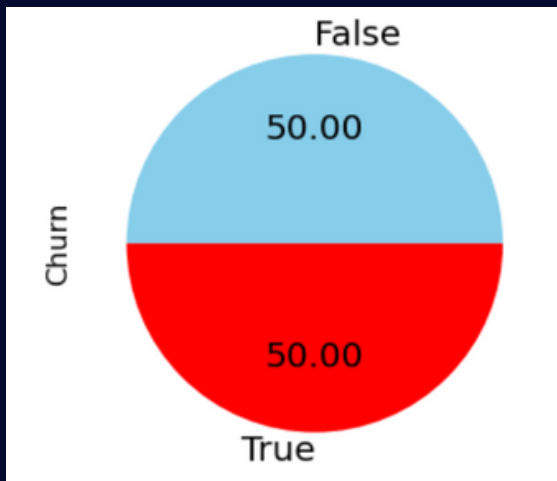
Achieving accurate and reliable results is crucial in any project. In order to achieve this, it's important to ensure that the features are appropriately scaled and normalized. There is a clear difference in scaling across the features, and a few of them are not normally distributed. Therefore, it's imperative to scale and normalize the features to ensure that the results are consistent and reliable.



Most features show a very low correlation.
However, there is a perfect positive link between total charge and total minutes over time. This is to be expected given that the charge for a call is determined by its length in minutes.

Total day minutes, total day charge, and customer service calls all show a weak positive relationship with churn. The other features show a minor connection with churn, around zero.

DATA PREPARATION FOR MACHINE LEARNING



Columns displaying total charge at different times are removed to address multicollinearity in features.

Train-test split: The data is separated into training and testing sets.

Dummy variables are constructed for categorical features.

SMOTE: SMOTE is used to address class imbalance issues by oversampling the minority class with replacement.

The pie chart below depicts the distribution of the target variable after applying SMOTE.

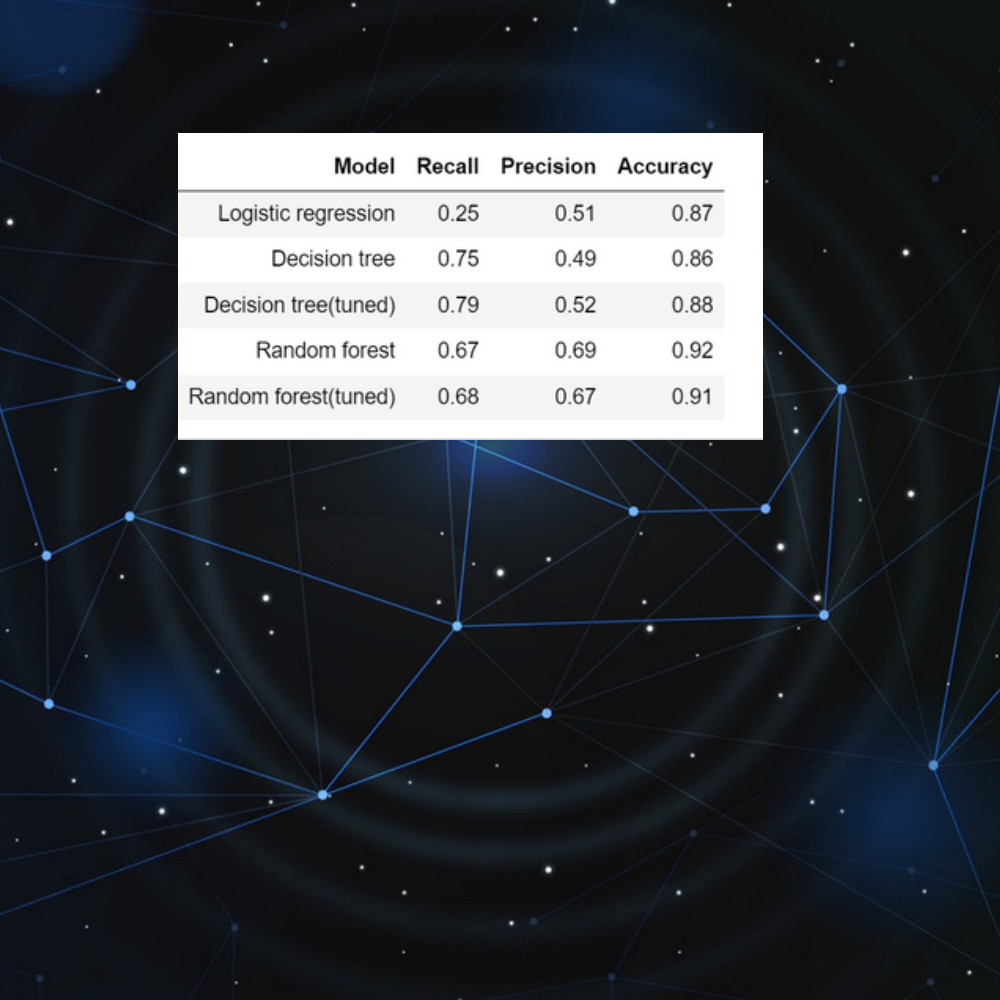
MODELING

Various models are built utilising logistic regression, decision tree, and random forest techniques.

Hyperparameter tuning is used with decision tree and random forest algorithms.

A pipeline is used to keep data from leaking. The pipeline scales data.

The following picture gives a summary of the models and their performance.



Model	Recall	Precision	Accuracy
Logistic regression	0.25	0.51	0.87
Decision tree	0.75	0.49	0.86
Decision tree(tuned)	0.79	0.52	0.88
Random forest	0.67	0.69	0.92
Random forest(tuned)	0.68	0.67	0.91

Logistic Regression: Has the lowest recall (0.25) but a relatively high accuracy (0.87). This suggests that while it correctly identifies 87% of all cases, it only identifies 25% of the actual positive cases. Its precision is moderate at 0.51.

Decision Tree: Shows a significant improvement in recall (0.75) compared to logistic regression, but its precision is lower (0.49), and accuracy is slightly less (0.86). This means it's better at identifying actual positives but at the cost of more false positives.

Decision Tree (tuned): Indicates a tuned or optimized version of the decision tree model. Tuning has improved both recall (0.79) and precision (0.52), as well as accuracy (0.88), suggesting that the model adjustments were beneficial.

Random Forest: This model has a lower recall (0.67) than the decision tree but a higher precision (0.69), and the highest accuracy (0.92) among the models before tuning. This suggests a good balance between recall and precision and overall the most accurate predictions.

Random Forest (tuned): The tuned version of the random forest has a slightly higher recall (0.68) but a slightly lower precision (0.67) and accuracy (0.91) compared to its untuned counterpart. The tuning in this case did not improve the model by a significant margin.

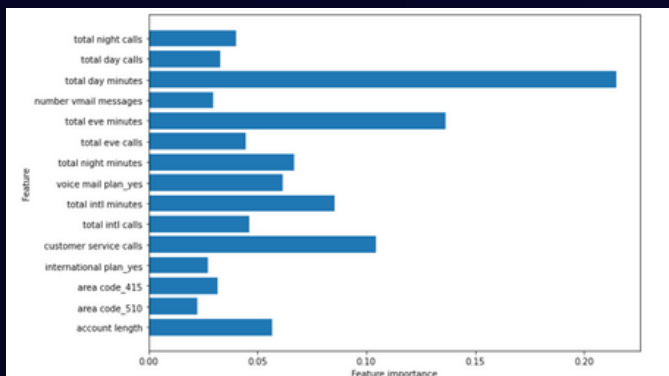
In summary, the table shows that the random forest model, especially the untuned version, seems to offer the best overall performance in terms of accuracy, while the tuned decision tree has the best recall. Precision is highest for the untuned random forest model. The choice of model would depend on what's more important for the specific application: for instance, if it's crucial to capture as many positives as possible, a model with higher recall would be preferred.

EVALUATION

The highest-performing model is a decision tree with tuned hyperparameters. It has the greatest recall score. The accuracy and precision ratings are higher than normal.

However, the recall score achieved is lower than the target of at least 85%.

The importance of each attribute in the best model is illustrated below.



The model with tuned hyperparameters excels in performance with the highest recall score, above-average accuracy and precision. However, its recall score falls below the 85% target. The feature importance of this top model is provided below

CONCLUSIONS AND RECOMMENDATIONS

The final model for predicting customer turnover is a decision tree with tuned hyperparameters. This model produces the fewest number of false negatives. The most essential features for predicting client churn include:

- Total day minutes: the total amount of minutes the consumer has spent on calls during the day.
- Total evening minutes: the total amount of minutes the consumer has spent on calls during the evening.
- customer service calls: number of calls made by the customer to customer service.
- Total international minutes: the total number of minutes the user has spent on overseas calls. Syriatel should provide good customer service in order to meet customers' expectations and analyse their interactions. They can then follow up on any positive or negative feedback received.



*Thank
you!*