taingocbui /
phase3_project

Code Issues Pull requests Actions Projects Wiki Security Insights Settir

0 stars **0** forks **1** watching **1** Branch **0** Tags Activity

Public repository

1 Branch 0 Tags Go to file t Go to file Add file Code

taingocbui Update final version b2993bd · 1 minute ago

data	include data	3 weeks ago
photos	Update final version	1 minute ago
.gitignore	update presentation.pdf	last week
README.md	Update final version	1 minute ago
notebook.pdf	Update final version	1 minute ago
presentation.pdf	Update final version	1 minute ago
project3.ipynb	Update final version	1 minute ago

README

SyriaTel Classification Model

Author: Tai Ngoc Bui
Date Completion: June 20th, 2024

1.Business Understanding

This project focuses on developing a robust and accurate predictive model to classify whether a customer will "soon" stop doing business with SyriaTel. This model should use recall rate as the main metric to ensure the model effectively identifies the most potential churners. The reason I use recall as the main metric here is Syria's business case should avoid minimize the number of customers who decide to churn but mistakenly labelled by the model as non-churners.

Besides building an effective model, this project also recommend the most important features to the SyrialTel team. These features will help stakeholders to implement targeted retention strategies to reduce customer attrition, thereby enhancing customer satisfaction and loyalty, and ultimately increasing revenue.

2.Data Understanding

This public dataset is provided by the Kaggle community's Churn in Telecom's dataset. It contains 20 predictor variables mostly about customer usage patterns. There are 3333 records in this dataset, out of which 483 customers are churners and the remaining 2850 are non-churners. Thus, the ratio of churners in this dataset is 14%. The data is clean overall with no duplicated row or Nan values.

Summary of Features in the Dataset

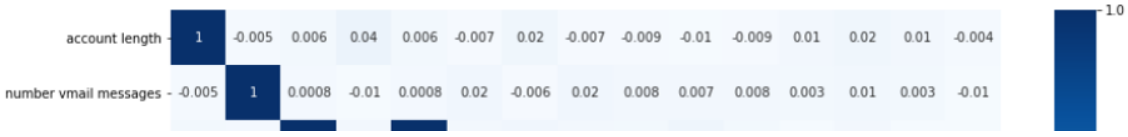
- state: the state the customer lives in
- account length: the number of days the customer has had an account
- area code: the area code of the customer
- phone number: the phone number of the customer
- international plan: true if the customer has the international plan, otherwise false
- voice mail plan: true if the customer has the voice mail plan, otherwise false
- number vmail messages: the number of voicemails the customer has sent
- total day minutes: total number of minutes the customer has been in calls during the day
- total day calls: total number of calls the user has done during the day
- total day charge: total amount of money the customer was charged by the Telecom company for calls during the day
- total eve minutes: total number of minutes the customer has been in calls during the evening
- total eve calls: total number of calls the customer has done during the evening
- total eve charge: total amount of money the customer was charged by the Telecom company for calls during the evening
- total night minutes: total number of minutes the customer has been in calls during the night
- total night calls: total number of calls the customer has done during the night
- total night charge: total amount of money the customer was charged by the Telecom company for calls during the night
- total intl minutes: total number of minutes the user has been in international calls
- total intl calls: total number of international calls the customer has done
- total intl charge: total amount of money the customer was charged by the Telecom company for international calls
- customer service calls: number of calls the customer has made to customer service
- churn: true if the customer terminated their contract, otherwise false

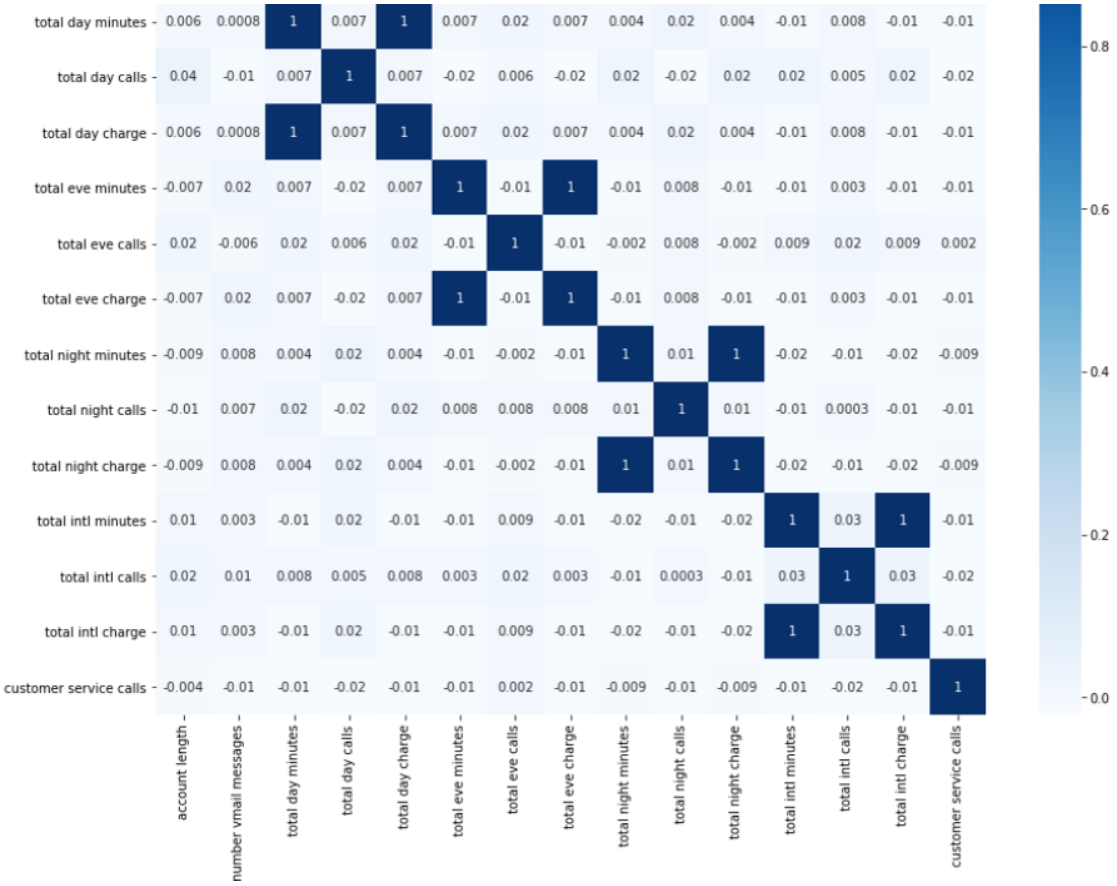
3. Objectives

- Develop a binary classification model to predict whether a customer will "soon" stop doing business with SyriaTel
- Explore and analyze to identify main features that would impact customers' decisions to stop using SyriaTel service
- Provide data-driven strategies to minimize the churning rate at SyriaTel.

5. Data Cleaning

This dataset has no Nan values and duplicated values. However, several features do have perfect correlations. Therefore, I decieded to drop several features which share perfect correlations to eliminate multicollinearity among the features.

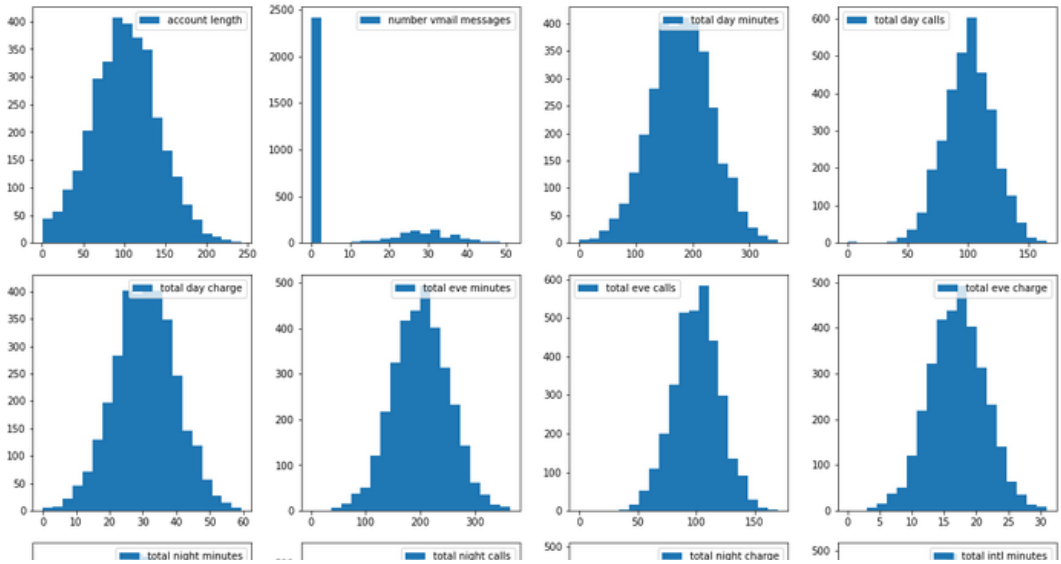


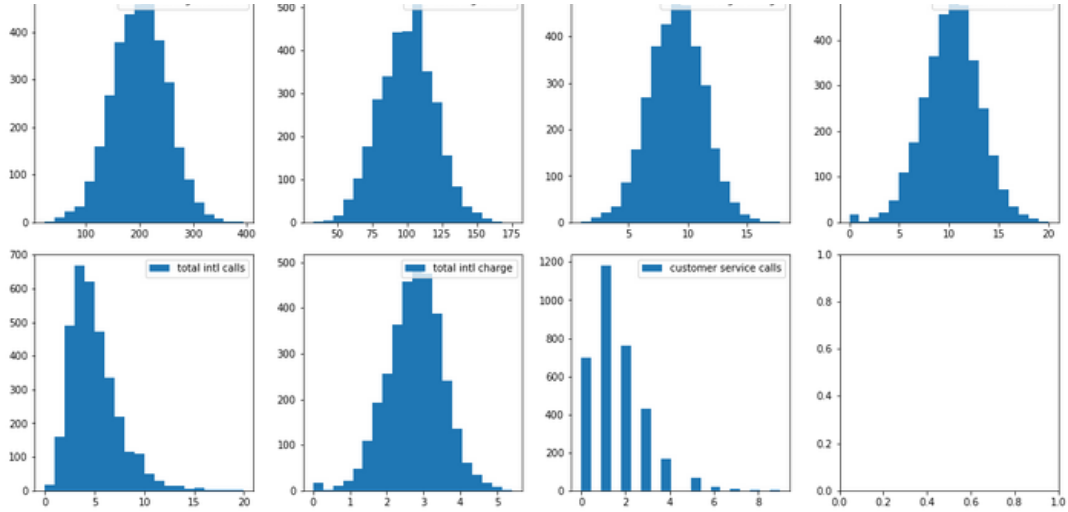


6. Preprocessing Process

Preprocessing process includes 4 steps:

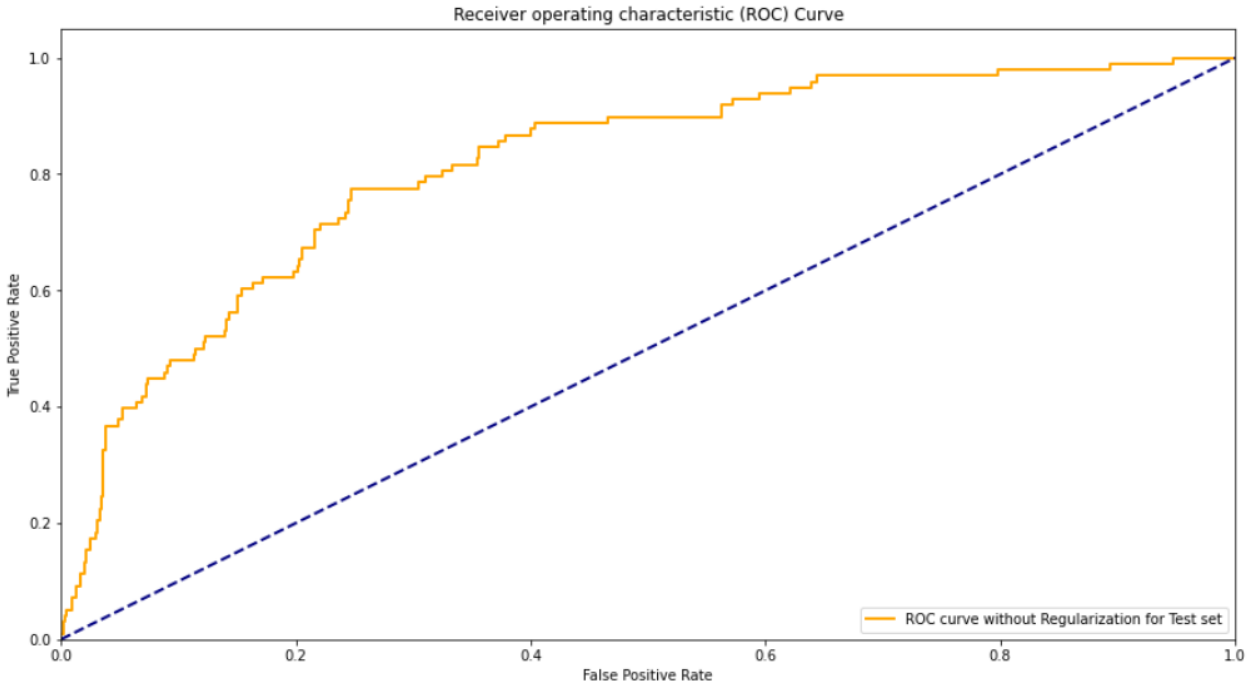
- Process categorical features as a one-hot numeric array using scikit-learn library
- Splitting dataset after one-hot encoded into train and test set with test set's size is 25% of dataset
- As most features are normally distributed, they will be normalized with Standard Scaler so that each numerical feature will have mean = 0 and standard deviation = 1.
- Finally, SMOTE is used to reduce data imbalance in the target.





7. Logistic Regression

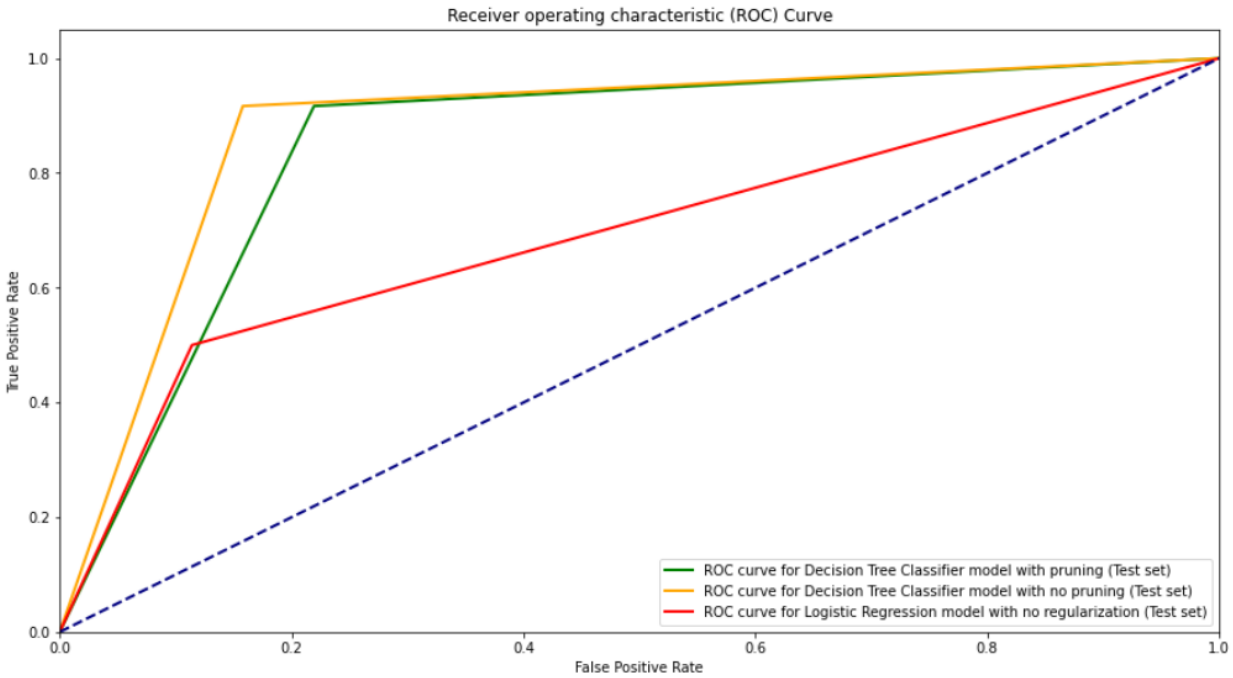
Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature. It is most commonly used when the data in question has binary output. My base logistic model without any regularization achieves only 44.89% recall rate on the test set. The Area under Curve for ROC is 0.81. Due to its significant low recall rate for both train and test set, no further tuning for this model needed.



8. Decision Tree Classifier

Decision Tree is a Supervised technique that can be used for both clasissification and regression problems, but is mostly preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

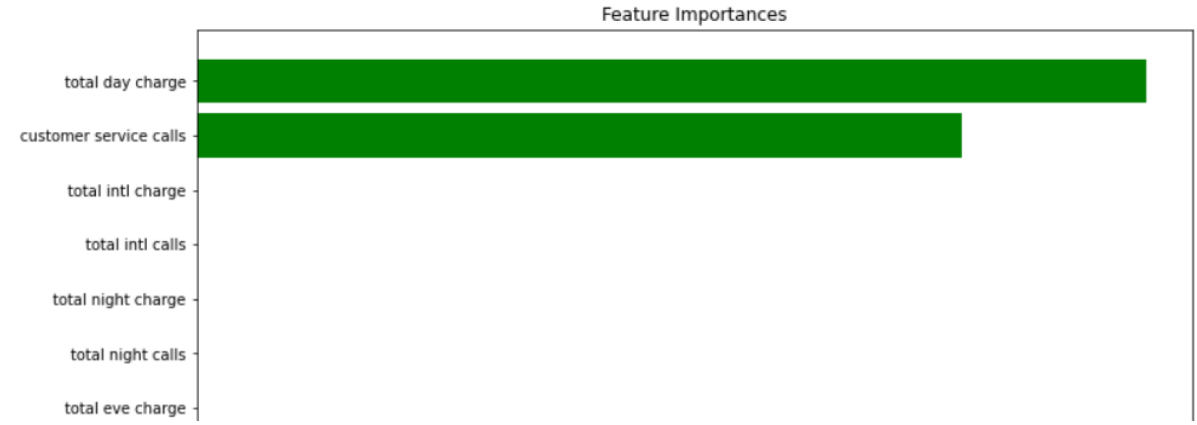
My base decision tree model resulted with 78.57% recall rate for test set. However, the model did indicated signs of overfitting due to its 100% recall rate for training set. To address overfitting problem, I tried to regulate the model by introducing new parameters such as max_depth, minimum sample split, or minimum sample leafs. The final decision tree model achieved a 87.13% recall rate for train set, 83.67% for test set and 91.6% for validation set. This final model is pruned with maximum depth set at 3 and minimum samples split set at 0.2. However, the trade off for eliminating overfitting is a lower ROC AUC (0.84 vs. 0.8793 in the base decision tree model)

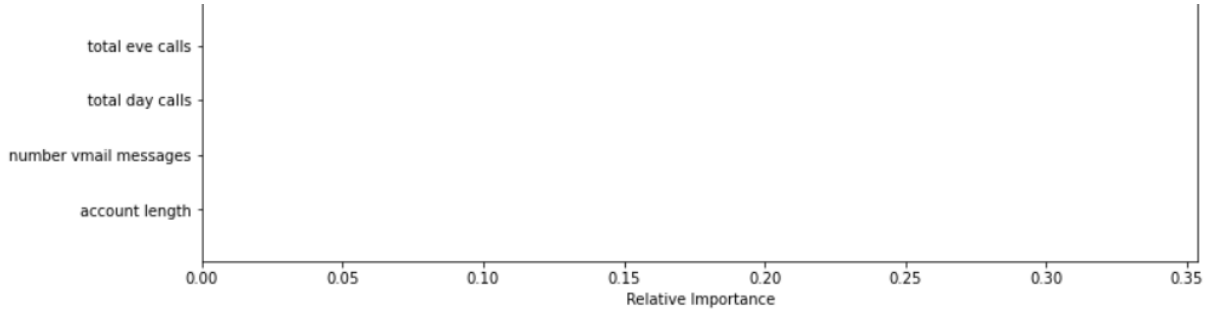


9. Conclusion

Based on recal rate of test set and validation set, I recommend Decision Tree classifier with maximum depth set = 3 and minimum samples split = 0.2 as the final classification model to SyriaTel stakeholders. This final model achieves my set target of maximizing recall rate and is sensitive to false negative rate.

On the other hand, the model also list the top 2 most important features including total day charge, and customer service calls. Based on these top features, I recommend SyriaTel to focus strategies on reduce charges to customer. Launching certain campaigns with discouts to users will greatly help SyriaTel reduce its churning rate. Furthermore, expanding customer service team to reduce customers' waiting time as well as offer training sessions to customer service team will have a great impact on SyriaTel.





8. Future Works

To better improve the quality of this report, I will extend this project using Random Forest Classifier model and Support Vector Machine model to possibly achieve a better accuracy and recall rate.

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

- **Jupyter Notebook** 100.0%