Customer Churn Prediction Model

Abstract

Several factors influence customer behavior resulting in churn such as corporations competition, changes in needs, etc. Companies are finding methods to ensure that customers remain loyal by recognizing churn customers and improving on things they lack at so that the customers stay. Moreover, a business lifecycle analysis requires looking at churn/retention rates to understand the health of its product. Therefore, to effectively manage customer churn, it is very essential to build an effective and accurate customer churn model.We aim to come up with a general analysis and modeling approach to predict churn which could very well be applied to various other datasets.

(I) Project Proposal

Research Goal

In today's businesses, it is more relevant to satisfy existing customers than to attract new customers who are characterized by a high attrition rate. Hence, companies have shifted their focus to existing customers more than onboarding the new ones. In fact, businesses increase their profit by decreasing churn customers. The definition of churn differs across businesses and the datasets can be highly domain dependent, the common attributes capture the level of engagement of a customer to the product like the number of complaints, visits to the dashboard and many more. With the analysis of dataset of users and its transactions, we seek to derive a general analysis and modeling approach to predict churn which could be utilized for various other datasets.

Questions we seek to Address

1) Does the performance of our models align with what was expected out of the Data Visualization Step?
2) Cross model performance comparison (across different classifiers)
3) What's the effect of data preparation and cleaning on model's performance with respect to the data in hand?

Methodology

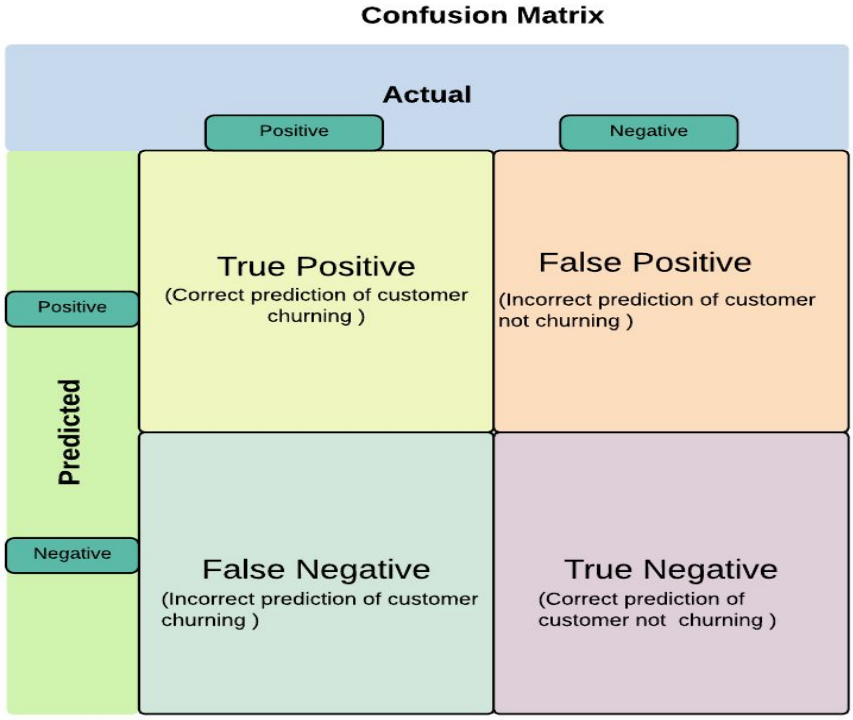
In this project, we will be first cleaning the data which involves removing duplicate values, missing values and resolving the inconsistencies in data . Next important step will be feature selection after which we will proceed with the data exploration and data visualization. We will be able to comment about the distribution of our data and the most appropriate model to be used for correct churn prediction by the data visualization. Next, we would be obtaining decision boundaries drawn between being churn and not being churn customers using 5 classification models :
a) Logistic Regression

b) K-Nearest Neighbors(KNN)
c) Support Vector Machine (SVM)
d) Random Forest
e) Naive Bayes Classifier

After this we will be analysing our results from the above 5 classifiers and comparing their performance based on F-Measure. This will help us check whether our best performance model aligns with our predicted model. We would also be showing the variation in results without data cleaning and with data cleaning in order to prove the importance of data preparation and analysis.

Metrics to measure Analysis Results :

We will be mainly focusing on F-Measure metric. We will be giving equal weightage to False Positive and False Negative and analysing the result. It will give us a good analysis of how the importance of False Positive and False Negative changes with respect to requirement of business. Then we will be weighing the False Negative more than False Positive and analysing the result. So intuitively, if our model incorrectly predicts ‘no churn’ for a customer (False Negative), then it can prove quite unprofitable than incorrectly predicting ‘churn’ for a customer (False Positive). In our case, False Negative will bring more loss to the business as it leads to missing focus on churn customers.



(II) Project Outline

Literature Review and Related Work :

http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html
https://gerardnico.com/data_mining/boundary
https://shapeofdata.wordpress.com/2013/07/02/decision-trees/

While the dataset has been picked up from a Kaggle data challenge and various competitors have tried to predict using the given dataset in problem, but as the challenges are more result driven, there are not much comparisons done across the choice of a model considered and the transformations taken. We seek to overcome this gap by analyzing the performances as affected by changing various parameters within a classifier like loss functions, kernels etc. as well as analyzing performances across different classifiers. Hence, we seek to emphasize the importance of Exploratory Data Analysis as a step in tandem with data modelling, which is often ignored in the haste of achieving the result.

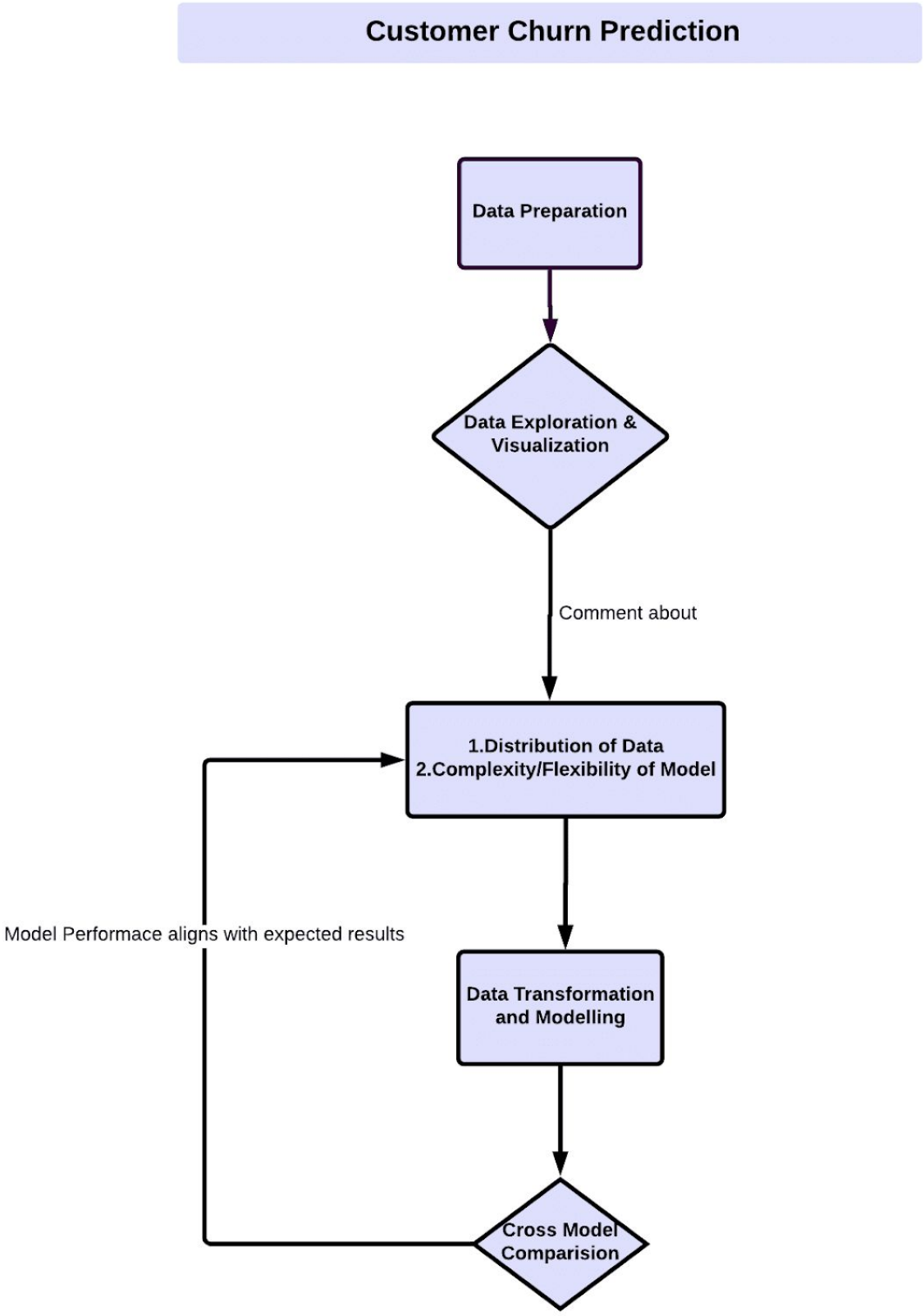
Data Source and Data description :

Dataset Type : Observational data picked up from the kaggle dataset as available here: https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data

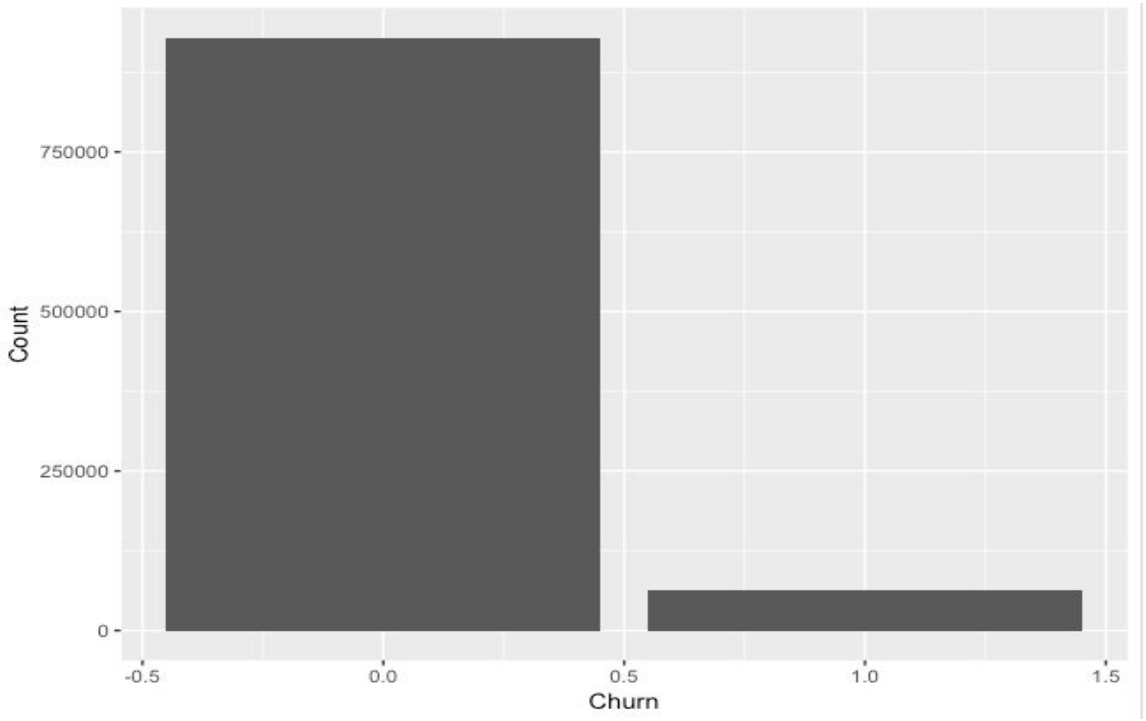
Feature Descriptions : The dataset is divided in several files which tracks the behaviors for 992,231 users. The behaviors tracked are the transactional behavior which has one to many mapping for user to transactions and contain around 21,547,746 records. The demographic data for the users which is a one on one mapping.

Dataset Size :
All in all, there are around 15 predictors for 992,231 different users.

Data processing and pipeline:



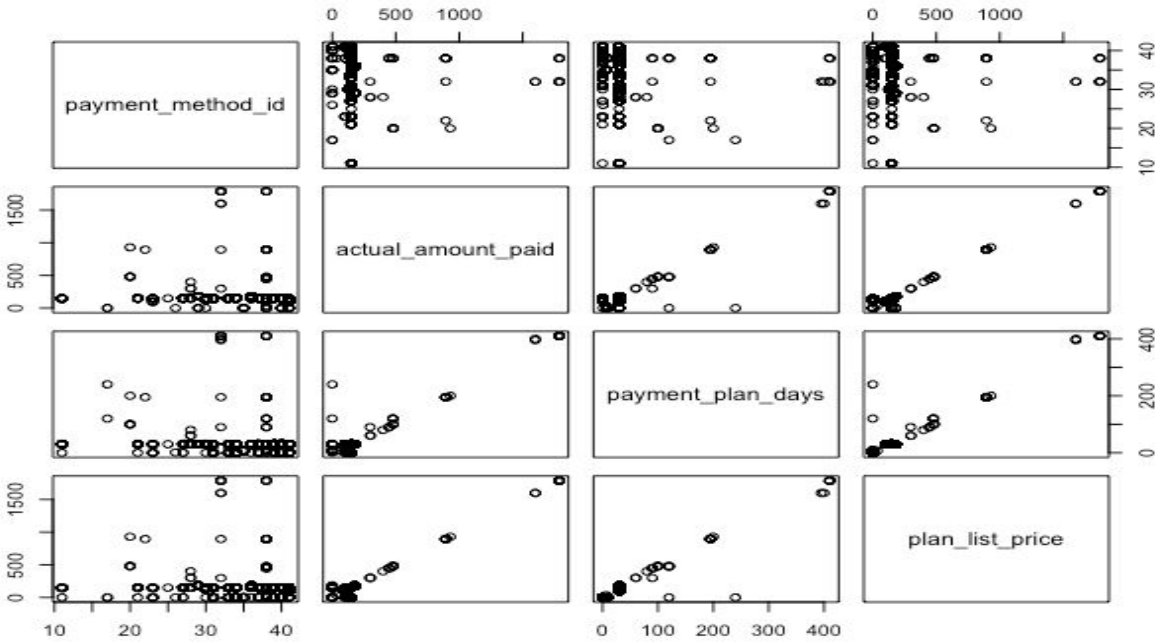
Data stylized facts:



Churn and not churn seem to be distributed in the following fashion in the dataset in study.

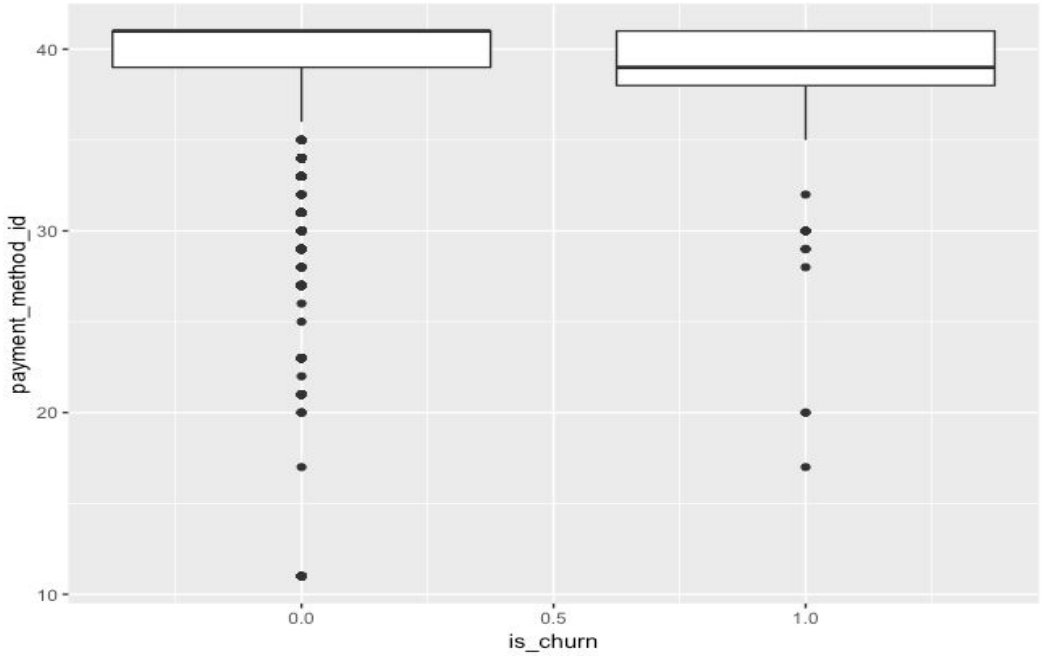
```
> trans_sam[duplicated(trans_sam),]
                                          msno payment_method_id payment_plan_days
14382954 mn1qFckfQ5GBltz0yuxIr3mIoAdKtLLkfZD1lrZLWjM=                41                30
         plan_list_price actual_amount_paid is_auto_renew transaction_date membership_expire_date
14382954             149                149             1         20151225               20141002
         is_cancel
14382954         1
```

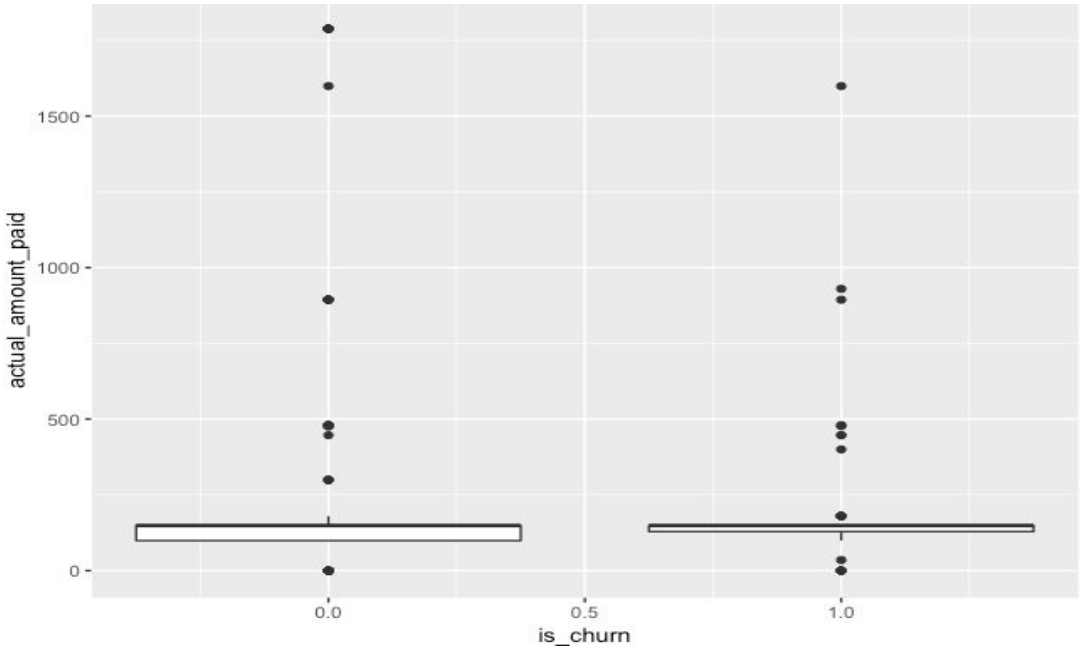
From the plot above, There lies little traces of duplicate data entries in our dataset. Though, the duplicate elements are very minimal as compared to the number of observations, it's still better to remove these duplicates rather than ignoring it.

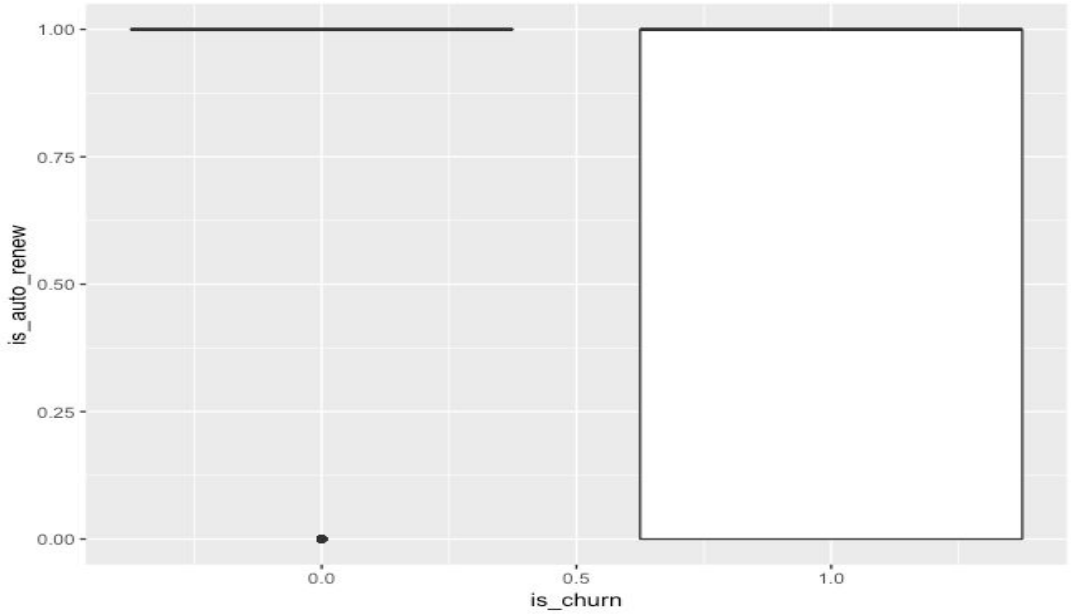


From the plot below, we can clearly observe the presence of predictors which are highly correlated to each other and needs to be taken care of before modelling or else, this would affect the model fit. The study also additionally aims to analyze on how not taking care of such cases affect the model.

Boxplot Visualization for class distribution and behavior analysis:







As we can see from the above plotted boxplots, while there are some significantly different behaviors across different classes in case of some parameters like renewing the subscription plan, but there are parameters/ dimensions where the behaviors overlap like payment method choice and actual amount paid towards the subscription.

We seek to develop a model which would help in segregating and correctly predicting the customers who are likely to churn.

Model selection:
We have planned to use 5 classifiers :

a) Logistic Regression,
b) K-Nearest Neighbors(KNN),
c) Support Vector Machine (SVM),
d) Random Forest and
e) Naive Bayes Classifier
as models for churn prediction for the WSDM dataset to compare results obtained from each and find which is better model to be used for churn prediction which can aptly draw decision boundary to divide data of users between to be churned and not . The better our model performs, more beneficial it would be for an organization with the perspective of business intelligence to take actions to focus the predicted users to be churned to entice and pull to section of loyal customers.

Tools :

Programming Languages : R , Python
Software Packages : R, Scikit-Learn
Applications : Rstudio