WSDM - KKBox's Churn Prediction Challenge

Can you predict when subscribers will churn?

CS 6375 Machine Learning Project Report

Ajay Attuchirayil Krishnankutty-	axa171830
Madhumitha Shankar-	mxs162530
Mohammed Amaan Shariff Shaik	mxs162030

April 30, 2018

CONTENTS

I INTRODUCTION ii

II BACKGROUND ii
- II-A ROC Curve. . . . ii
- II-B Precision and Recall ii
- II-C Class Imbalance ii

III PROBLEM DESCRIPTION AND ALGORITHM iii
- III-A Data Description iii
  - III-A.1 Train.csv iii
  - III-A.2 Transactions.csv iii
  - III-A.3 Member.csv iii
  - III-A.4 User logs.csv iii
- III-B Algorithm Definition iv

IV PRE-PROCESSING TECHNIQUES iv
- IV-A Data Pre-Processing iv
  - IV-A.1 Memory Reduction iv
  - IV-A.2 Pre-Processing of Members. . . . iv

V FEATURE ENGINEERING. . . . v

VI EXPERIMENTAL EVALUATION. . . . v
- VI-A Grouping Data v
- VI-B Merging Members and Transactions. . . . v
- VI-C Merging Data Frame with Train and Test for Analysis. . . . vi
- VI-D Under Sampling on Data Frame vi
- VI-E Modelling vi
- VI-F Grid Search, Cross Validation and Random Forest. . . . vi

VII RESULTS vi
- VII-A Precision- Recall vi
- VII-B ROC curve vi

VIII FUTURE WORK vii

IX CONCLUSION vii

References vii

LIST OF FIGURES

Fig 1: Users vs Churn. ii

Fig 2: Features of Transaction dataset. ii

Fig 3: Features of Members dataset. iii

Fig 4: Gender of the Users. iii

Fig 5: Data frames After Feature Engineering. v

Fig 6: Heatmap of pairwise correlation between features. v

Fig 7: Feature Importance. v

Fig 8: Precision- Recall-F1-score-Support. vi

Fig 9: ROC curve. vi

I. INTRODUCTION

In a subscription business, accurate prediction of upgradation/continuing of existing subscription is a critical prospective in terms of long term acumen and profitability for the company. The slightest variation will affect a commercial profit drastically. WSDM KKBox's Churn Prediction Challenge will help in increase the accuracy in predictability of this business. This challenge helps us to find an alternative/improvement from the current technique (Survival analysis technique).

The project aims at formulating an algorithm which will help in determining whether a user leaves/quits a subscription, helping KKBOX to maintain the subscribed users. Details of each user behavior is available on the website and further, analysis of the behavior is to be done and correlate the reasons for users to churn or continue the subscription.

The KKBOX also included the users behavior so that we can explore different users behavior outside the train and test sets. Since it is a classification problem, we choose to use technique likes ensemble methods and Neural Networks to execute the results. This will let the learner test the model and evaluate the result using various scoring parameters and metric evaluations. Machine Learning libraries like Scikit Learn, open source tool Panda and numpy packages will be used to develop the project. We will build the algorithm and predict whether a subscription user will churn using a dataset from KKBOX.33.

II. BACKGROUND

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes

- NumPy: Base n-dimensional array package.
- SciPy: Fundamental library for scientific computing.
- Matplotlib: Comprehensive 2D/3D plotting
- I Python: Enhanced interactive console.
- Pandas: Data structures and analysis.

Some popular groups of models provided by scikit-learn include

1) Cross Validation: for estimating the performance of supervised models on unseen data.
2) Dimensionality Reduction: for reducing the number of attributes in data for summarization, visualization, and feature selection such as Principal component analysis.
3) Ensemble methods: for combining the predictions of multiple supervised models.
4) Feature extraction: for defining attributes in image and text data.
5) Feature selection: for identifying meaningful attributes from which to create supervised models.
6) Parameter Tuning: for getting the most out of supervised models.
7) Supervised Models: a vast array not limited to generalized linear models, Discriminate Analysis, Naive Bayes, Lazy Methods, Neural Networks, Support Vector Machines, and Decision Trees.

We tend to use accuracy because everyone has an idea of what it means rather than because it is the best tool for the task. Accuracy does not provide a useful assessment on several crucial problems, often we need to go beyond accuracy when developing classification models. We can use Recall, precision, and the ROC curve.

A. ROC Curve

The ROC curve is a fundamental tool for diagnostic test evaluation. In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between the two groups. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore, the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test.

B. Precision and Recall

Recall and Precision are better suited metrics for imbalanced classification tasks. It can be said that precision and recall are totally dependent on the factor of relevance. Recall expresses the ability of a classification model to identify all relevant instances in a dataset whereas precision expresses the proportion of the data points our model says was relevant actually were relevant. To fully evaluate the effectiveness of a model, you must examine both precision and recall. Unfortunately, precision and recall are often in tension. That is, improving precision typically reduces recall and vice versa.

C. Class Imbalance

Class Imbalance problem in machine learning where the total number of a class of data (positive) is far less than the

total number of another class of data (negative). In many cases when provided with such kind of data it leads to false predictions and false accuracy. The different ways to tackle the imbalanced data is cost function-based approach, sampling based approach. Sampling based can be broken into three major categories: over sampling, under sampling and hybrid of over sampling and under sampling.

III. PROBLEM DESCRIPTION AND ALGORITHM

A. Data Description

The Xbox music service provider, contains a list of users having monthly subscription plan. The subscriber can choose to renew or cancel the subscription plan after the 30 days from the day of joining. The member can also choose to cancel the auto-renewal plan. The dataset for the company is provided in five different files. The target features of this dataset are transaction date, membership expiration date, and is_cancel. The broad description of dataset and their features are as follows:

1) Train.csv : It consists the data of the users whose membership would expire in one months time. The month of February is taken into consideration. The month of March of the year 2017 is considered for the prediction of the churn renewal. The train data set consists of the following target variables: user id and is churn.

From the fig1, we can see that only 60000 people have

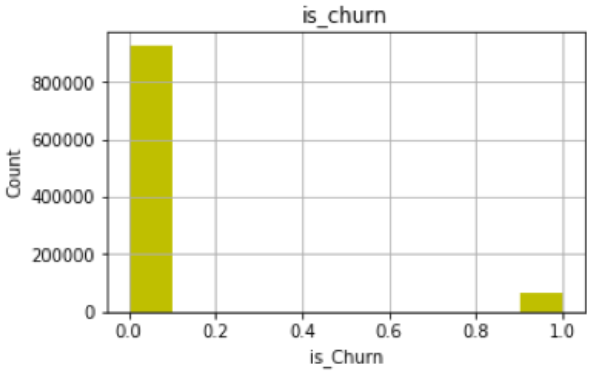

Fig. 1. Users vs Churn

churned so far, and we need to find out when will the remaining people churn.

2) Transactions.csv : This consists of information of all the transaction the user has completed up until February 2017. The features present in this data set are msno, payment_method_id, payment_plan_days, plan_list_price, actual_ amount_paid, is_auto_renew, transaction_date ,membership_expire_date and is_cancel.

The above figure is a plot of all the features of the transaction dataset. The feature named as is_auto_renew has binary values. Popularly, we can say that users renew their subscription. To know whether a user has canceled the

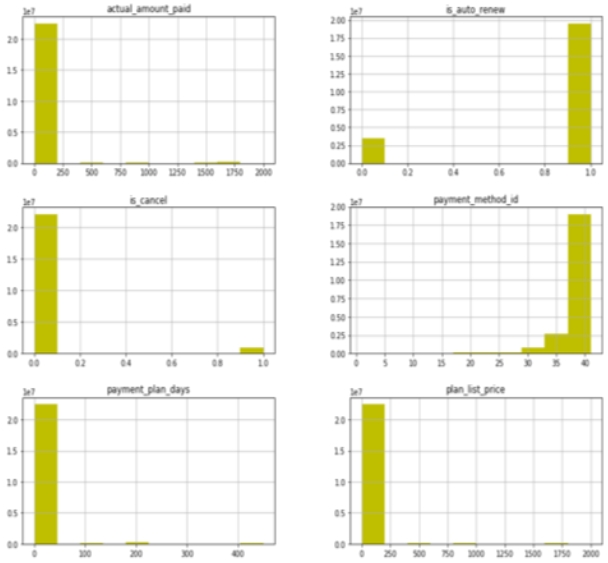

Fig. 2. Features of Transaction dataset

subscription or not is determined by the feature is_cancel. The difference between the features transaction_date and membership_expire_date gives period a user has been on the website. This information provides whether the person churned or not.

In the feature named payment_method_id the most frequently used ID number is payment ID 41. The feature payment_plan_days column, shows that the value is 450 days if the user has opted for either manual or automatic resubscription.

3) Member.csv : This consists of all the information of the subscription services taken by the users. The columns of the dataset are msno, city, bd, gender, registered_via. registration_init_time, and expiration_date. The figure 3 shows the plots of different features in the member dataset.

The feature named bd (birthday date) provides information regarding the age of a member. The feature named city is encoded with data where in 21 cities are included.

The feature named registered_via has data in the encoded format. The registration_init_time is a date encoded as an integer. To use this information, it needs to be decoded to date format.

The above dataset (in figure 4) pot shows the features gender, where it gives the information is the user has provided the gender as male, female or not given. The feature provides the data in categorical format.

4) User_logs.csv : This another dataset containing the data of users behavioral characteristic information. This dataset has information from the month of Jan 2015 to Feb 2017. The columns of the dataset are msno, date, num_25, num_50, num_75, num_985, num_100, num_unq and total_secs. The feature total_secs contains the total number of seconds the user played songs on a particular day.

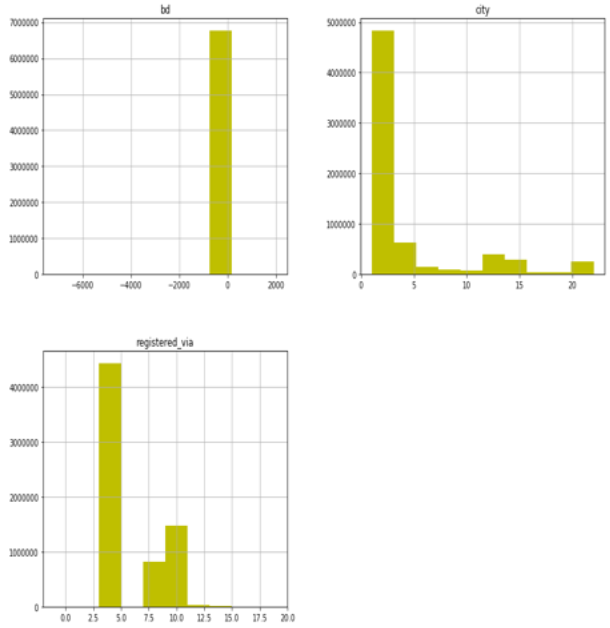

Fig. 3. Features of Transaction dataset

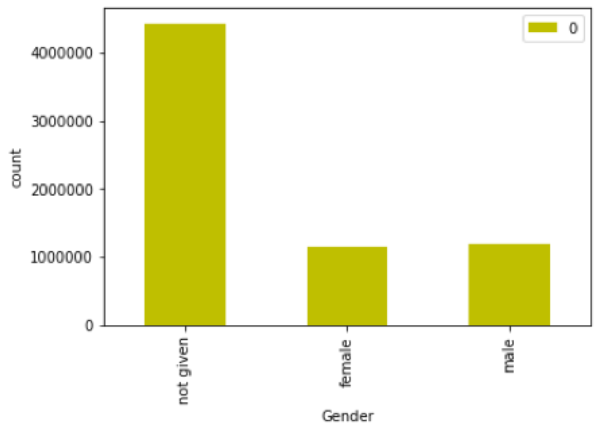

Fig. 4. Gender of the users

The feature num_25 has the number of songs played for less than 25% of the time frame. The feature num_50 has the total number of songs that were played between 25% and 50% of the songs length.
The features num_unq is the key feature containing number of unique songs the user played on that day.
Test.csv contains the same dataset as in train. The two main target features are msno and is_churn_column.

B. Algorithm Definition

The churn prediction challenge has a diverse dataset and there are many aspects that need to be considered for developing a model. The data set is vast and contains millions of information. The Since it is a classification problem, we choose to use technique likes ensemble methods and Neural Networks to execute the results. The algorithm that we choose to for the classification problem considers all the aspects of the dataset. There are various techniques considered to reduce the memory size of the data before pre-processing it. We plan to engineer new features from the existing ones which can be more meaningful. One of the derived feature is taking the existing features the price and paid amount to calculate the discount. Later, based on the ID of the user the different datasets are combined and then this is combined with the test

And the train data. The best feature is described from the pre-processed data using based on Correlation. Various models like Random Forest and Artificial Neural Networks are executed on this data. ROC curve, precision call must be performed on the resultant data for choosing the best model based on the evaluations.

IV. Pre-processing techniques

A. Data Pre-Processing

The dataset available here is very large, and to extract some meaningful features from it we used some pre-processing techniques. Missing values in various features were handled by either removing them if null was the majority for that feature or by imputing them with their mean or median values. Then to reduce memory usage while training we updated the data types of many features.

1) Memory Reduction: In this we parse through the whole set of numeric features and change the reduce the data type if required based on the Min and Max value of each feature. For instance, if the Min and Max value lies in the range -128 to 127 then in that case datatype can be updated to int8 if its something larger than that. Similarly, we convert datatype to int16 if the range is between 32767 to -32768, and int32 for the range 2147483647 to -2147483648.

2) Pre-Processing of Members:

- Birth Date / Age (bd) - The feature bd has lot of outliers present in the form of either negative values or age greater than 1000. So, we clip these outlying values keeping threshold values either greater than 10 or less than 100. So now we have ages in a more practical range, we now convert this feature to categorical with the use of binning. We use a Label Encoder and convert this feature into following bins = Age Group: 0-18 (Children), 18-30 (Teenagers), 30-50 (Medium), 50-100 (Old). From this binning we get an information that majority users come from the teenager section of age.

- Gender - Majority of the users have not specified their gender, but of those who have mentioned there is a similar ratio among Male and Female. We convert this feature to numerical one by updating Males as 1 and Females as 2 and the ones that did not specify as 0.

V. Feature Engineering

We will now try to bring out some new features based on the existing ones, which will help us better in the prediction process [Th data frame Transactions had multiple features which on analysis was found out could be engineered to more meaningful features.

1) Discount: We try to find out if any of the users received any kind of discounts on their subscription purchase. The same we can get by considering the difference between plan_list_price and actual_amount_paid both from the transactions data frame. We get values that are either positive or negative, so we convert the negatives to 0 showing no discount and 1 in the positive case where there is discount. We add this up to a new feature is_discount.
2) Amount Per Day: We now turn to see how much the users have to pay per day for the subscription plan chosen by them. We can find this out by dividing actual_amount_paid by the payment_plan_days. Again, we get some outliers in this in form of infinite values, we fill up those using 0.
3) Membership Duration: This was easy, we just found the difference between the membership expiration date and the transaction date, and then converted it to days. It had some negative values we clipped those off substituting by 0.
4) Auto Renew: We convert this feature to a binary feature having 0s and 1s based on if the user has Auto renewal option turned on. The feature auto renew can affect the churning nature of a user and it was calculated if the user auto-renewed or not in each case.
5) Memory reduction was executed on the data frame to reduce memory of the new frame with new features.

VI. Experimental Evaluation

A. Grouping Data

In the transactions data frame, we have multiple records for same users who may have done more than one transactions. Here, msno stands for the user id in the transaction data frame, so now we need to groupby all such users. We use some metrics from the transaction data preprocessing. We define a mean and mode function and apply it the features of the data frame, to generate following findings. The payment_method_id which has been used more by the user is assigned to that user. We assign the mean of payment_plan_days to the user because that can extrapolate the average usage of the user in the website every time he/she makes a payment. This is the average payment plan days used by the member for a given transaction. For the feature not_auto_renew the count of which a user has auto-renewed is calculated. It gives the number of times the user has not auto renewed his license. We get the frequency of each user performing a transaction is calculated using the count of msno which is the unique ID for each user. The feature is_cancel is gives the number of times the user has canceled transactions. Also, the feature is_discount shows the number of times user has received discounts from KKBox. The mean of the feature membership_duration gives the average membership duration of the user. Hence the final processed data frame Transactions contains each user ID and the following features:

1) payment_method_id : Most Frequent Payment method.
2) payment_plan_days: Mean of payment plan days the user used every time he made a transaction.
3) not_auto_renew: Number of times a user auto-renewed.
4) msno_count: Number of transactions performed by the user.
5) is_cancel: Number of times a transaction cancelled by the user.
6) is_discount: Number of times a discount was awarded to the user.
7) amount_per_day: Payment user has to pay per day based on the plan selected.
8) membership_duration: Average membership duration of the user.
9) membership_expire_date: Existing expiry of membership of a user.

B. Merging Members and Transactions

We now have these new features in the group by data frame. Now we merge it with the transactions data frame to get a better idea. Since there are users whose information has not been given, a left join of the frame transactions was performed with the frame members.

The gender features null values were filled with zeroes, categorical features bd, city and registered_via had their null values replaced with their respective modes. The numerical variant of bd, in this column the null values have been replaced with the mean.

A feature long_time_user was engineered by combining these two data frames. We get two features membership_expire_date (max of expiration date from the transactions data) and the registration_init_time(the time when the user has registered). Difference of these two columns gives the duration of time for which the user is subsribed with KKBox. If we divide with 365. we can get the number of years the user has been with KKBox.

Since main features have been extracted we can remove base features like 'membership_expire_date', 'registration_init_time' and 'reg_mem_duration which will help in reducing memory usage.

A minmax scalar is used the data points under features payment_plan_days', 'not_auto_renew', 'msno_ount', 'is_cancel', 'is_discount', 'amount_per_day, 'membership_duration', 'long_time_user' and 'bd' have been scaled to values between 0 and 1.

Also, all the categorical variables have been converted into numerical using label encoding. One hot encoding has not been used since there are many values for the categorical values. Therefore, one hot encoding results in having many columns in the final data frame. Categorical feature's

payment_method_id,'city','registered_via' are converted to numerical using label encoder. Hence the final data set after feature engineering and preprocessing is as below:

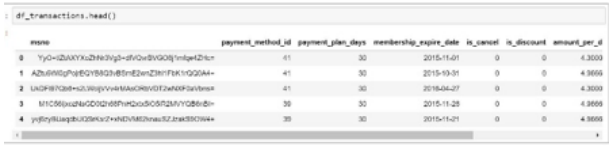

Fig. 5. Dataframe after feature engineering

C. Merging Data Frame with Train and Test for Analysis

Finally, we merge this processed data with the test and training data we already have based on the userID which corresponds to msno.
A correlation heat map between the features of this merged data frame is generated and shown below: From the correlation Heatmap it has been overserved that there is high negative correlation between is_churn and payment_id and msno_count. Also, there is a high positive correlation between is_churn and payment_plan_days, not_auto_renew, is_cancel and membership_duration.

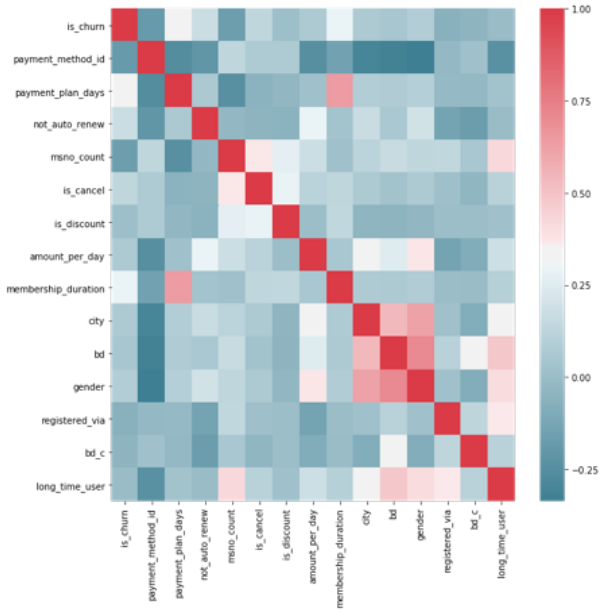

Fig. 6. Heatmap of pairwise correlation between features.

D. Under Sampling on Data Frame

From Fig 1 we can say that the data is unbalanced as Class 0 is close to five times that of Class i.e. every time almost one user churns when five users continue. So, we do under sampling, to reduce the class imbalance by taking random samples from train and test data.

E. Modelling

Several Modellings have been executed and results have been analyzed. The best Model turned out to be Random Forest and Grid Search was performed to analyze its best parameters.

F. Grid Search, Cross Validation and Random Forest

GridSearch as well Cross validation are used to find out the best parameters. After using grid search we plotted the feature importance bar graph as shown in the Fig 12. Ensemble methods such as random forest and Gradient boosting have been used due to class imbalance. The data was tested with various models and the best accuracy was evaluated. The best model is presented here. Some parameters have been given to Random forest classifier to identify the best one. The grid search returns the best parameters. This was performed several times and with different parameters.

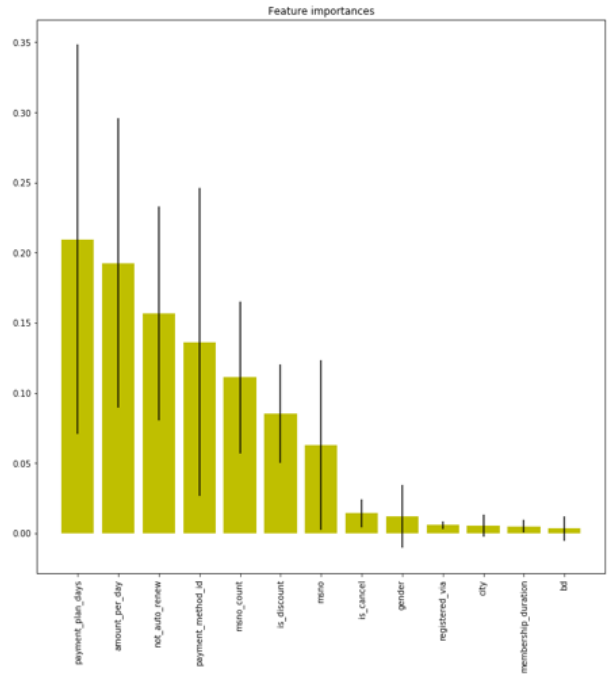

Fig. 7. Feature Importance

VII. Results

A. Precision- Recall

The Results have been evaluated using precision, Recall, F1 score and support.
Precision was found as 0.94 and recall 0.93. The F1-score value is 0.93.

B. ROC curve

ROC curve was plotted to analyze the results better. The closeness of the curve to the top left corner shows the performance of the prediction made by our model. The area under ROC curve for the best model is 0.95. The results have been satisfactory and in-line with the initial assumptions.

```
: print(classification_report(y_test,t_pred))
```

	precision	recall	f1-score	support
0	0.98	0.95	0.96	543775
1	0.53	0.71	0.61	45393
avg / total	0.94	0.93	0.93	589168

Fig. 8. Precision- Recall-F1-score-Support

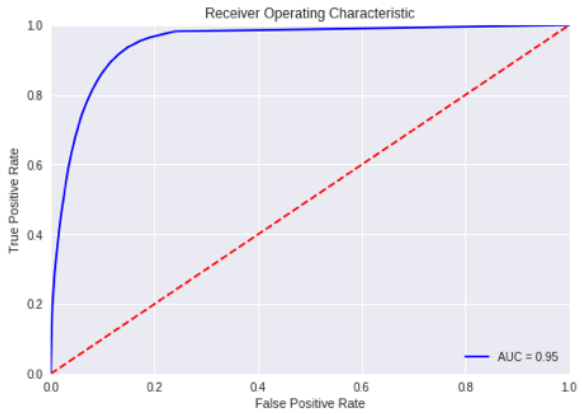

Fig. 9. Precision- Recall-F1-score-Support

VIII. Future Work

The normalization and scaling of the data of user_logs is one suggested area which can give us the information on the correlation of features and on an average basis a relation between listening habit and duration of time of a single song a user listens.to, on an average basis. Also we found out that the user_log contains a huge amount of null values for user behavioral data. If we had more information on this listening habits and behavioral data, we can develop a recommender system as well for the users, which is one of the interesting area.

IX. Conclusion

In the project we aimed at formulating an algorithm which will help in determining whether a user leaves/quits a subscription, helping KKBOX to maintain the subscribed users. The dataset for this project is obtained from the Kaggle challenge WSDM - KKBox's Churn Prediction Challenge. Initially, after performing data analysis, we concluded that the dataset set contains diverse features such as users gender, age, date of birth, payment month and method, cancellation history, log of the songs listened by the user, type of plan chosen by the user, registration method, transaction date etc. After the exploration of the data and applying pre-processing techniques, we derived new features from the existing feature which would help us develop a better model. The engineered features were tested on ensemble techniques like Random Forest, Gradient boosting, and Neural Networks. The best model suggested predicted the test data with a precision and recall of 0.94 and 0.93 respectively. From the ROC curve we see that AUC is 0.95 which is excellent for a prediction. By this we can conclude that the algorithms applied works with a very high accuracy and gives new insights to why users leave which is better than the conventional methods.

References

[1] Machine Learning, Tom Mitchell, McGraw Hill, 1997.
[2] http://pandas.pydata.org/pandas-docs/stable/
[3] https://www.kaggle.com/headsortails/should-i-stay-or-should-i-go-kkbox-eda
[4] https://www.kaggle.com/the1owl/regressing-during-insomnia-0-21496
[5] https://www.kaggle.com/jeru666/memory-reduction-and-data-insights
[6] https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/
[7] http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[8] http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
[9] http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[10] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html
[11] http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
[12] https://docs.scipy.org/doc/numpy/