BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

Assignment -10 (Music streaming service)

Group Number – 93, Mudit Jain, Rajesh Mohite, Vivek Singh,

Introduction to Data Science

M.Tech Data Science and Engineering – Cluster Batch 4

14-FEB-2021

Overview

- Objective –

Predict if subscription users of a music streaming service will churn or stay after their current membership expires.

Posing as Machine Learning Problem

Binary class classification: is_churn either 0 or 1.

- Methodology –

1. From Problem to Approach –
   Understanding the problem statement of to predict is a user will churn once its subscription expires.
2. From Requirements to Collection -
   Collect the relevant data i.e., the data link shared in the assignment didn't contained the churn status of any user. So we looked for another data set of the same music streaming service and extracted relevant information from it like – User info, subscription details and transaction details.
3. From Understanding to Preparation –
   Prepared the relevant data to predict but before that performing EDA on it and process the NULL or NAN fields and with relevant data, dropping any field which is not required. Also, performed Feature Engineering on the data.
4. From Modelling to Evaluation –
   Evaluate the performance of the two models used and find out which is performing better.
5. From Deployment to Feedback –
   Test the model on test data and verify its performance.

Methodology

- The 2 classifiers used –
  1. Logistic Regression
  2. Decision Tree
- Ensemble pipeline – We have not created ensemble pipeline since we have picked up only the relevant data i.e., train, transactions and subscription data. Also, our goal as given in the question was to perform prediction via Logisctic Regression and Decision Tree and compare their performance.
- Other models considered – Since the models mentioned in the question were – Logistic Regression and Decision Tree we didn't considered any other model. However, for finding out the top 10 features we did used – Linear Regression algo.
- Hyper-parameter tuning – No hyper parameter tuning was applied since we tested the Logistic Regression with – "GridSearchCV" but it resulted in the same score as without hyperparameter tuning.

Dataset

- How many features – We have considered 15 features, which after applying One-Hot Encoding on categorical data increased to 37
- Size of the dataset – 589MB
- Multiple files – 3 files considered (train_v2.csv, members_v3.csv and transactions_v2.csv)
- What kind of data – Numerical
- Balanced or imbalanced – The target feature's data distribution was – imbalanced.
- Distribution of Training set, validation set, testing set – Training data was splitted in 80-20 ratio between training and testing data.
- Missing data and Preprocessing challenges –
  1. We have applied "inner join" to merge the data since lot of other fields were not present in other joins. Also, important fields like – age, gender had missing data issues in them.
  2. Before preprocessing we have done EDA to identify the important or relevant features. Also, while in preprocessing we realized that data needs to merge from different data frames.
  3. Transforming data type of the data on which prediction needs to be done since prediction or binary classification can't happen on `object` types.

BITS Pilani
Pilani | Dubai | Goa | Hyderabad
Work Integrated Learning Programmes

Assignment -10 (Music streaming service)

Group Number – 93, Mudit Jain, Rajesh Mohite, Vivek Singh,
Introduction to Data Science
M.Tech Data Science and Engineering – Cluster Batch 4

14-FEB-2021

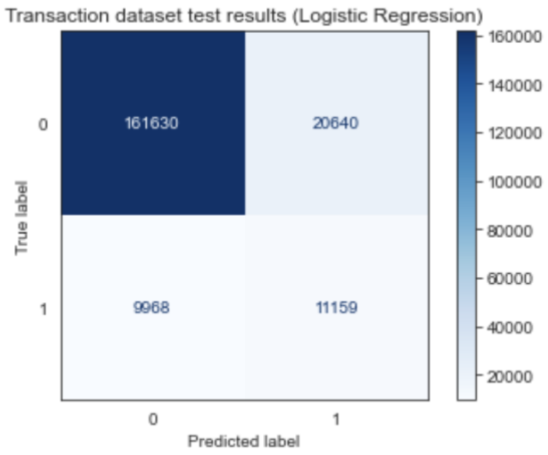
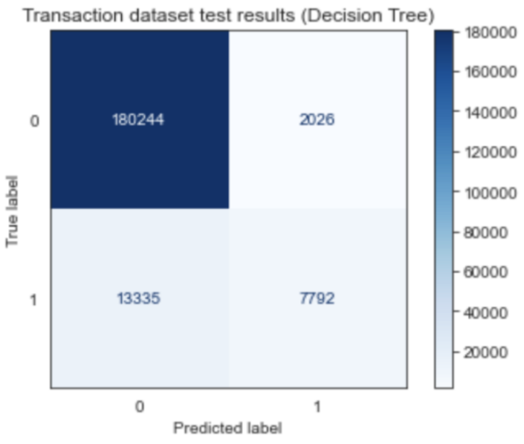
Feature Engineering Techniques

- Features removed –
  1. msno
  2. registration_init_time
  3. transaction_date
  4. membership_expire_date
- Feature creation –
  Not required to create any new feature however, we splitted the categorical features via One-hot Encoding technique.
- Feature ranking –

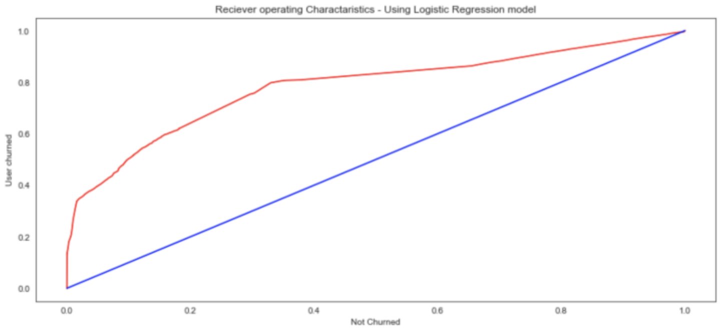
  Identified top 10 features as (in descending Order)-
  1. Feature 4 – actual_amount_paid
  2. Feature 5 – is_auto_renew
  3. Feature 6 – transaction_date
  4. Feature 1 – payment_method_id
  5. Feature 7 - membership_expire_date
  6. Feature 0 – msno
  7. Feature 2 - payment_plan_days
  8. Feature 3 - plan_list_price
  9. Feature 8 – is_cancel
  10. Feature 9 – city
- Class imbalance treatment – Not required
- Any other – Not required

Results

- Table for the evaluation metric for each ML technique used

Confusion Matrix for Decision Tree – (92% accuracy)	Confusion Matrix for Logistic Regression – (84.9% accuracy)
	

- Plot of the curves
  ROC Curve –



- Conclusion –
  1. The accuracy we achieved using Logistic regression is 84.9% and using Decision Tree is 92% respectively.
  2. The performance of Decision tree is more than compare to Logistic regression