CSP 554 – Big Data Technologies - Project Proposal

A20495939 - Jasleen Bhatia
A20473187 - Kranti Mahadev
A20490179 - Nandish Bhagat
A20460031 - Taranpreet Bhatia

Customer Churn Prediction on Sparkify Dataset

ABSTRACT

Sparkify is a music streaming dataset. Customer may switch from a free version to a premium membership or paid subscriptions, or vice versa as they interact with the music service. As we know, customer churn occurs when a customer chooses to terminate or degrade their membership to a certain plan. The aim of this project is to anticipate the customer churn on the real-world resemblance music steaming dataset.

RESEARCH GOAL

The project mainly focuses on predicting customer churn (detect if the specific customer will cancel the service) on the Sparkify Dataset (mini dataset – 128 Mb & full dataset – 12 Gb). The dataset is composed of numerous user events in the audio streaming service provider like Spotify and Pandora. At first, we will analyze the mini dataset using spark Dataframe, Spark SQL & Spark ML API on local using Pyspark and then deploy the spark cluster on AWS to run the model on full dataset. Also, we will use Tableau to visualize the insights from the dataset.

OBJECTIVE

To achieve the aim of this project, we will focus and investigate customers churn by gathering insights from the dataset -

- Distribution of each feature
- Distribution of customers by gender
- Churn by location/states
- Visualization of free vs paid customers
- Predicting customer churn.

PROJECT OUTLINE

Data Processing

- Find and remove empty values in all datasets.
- Change the data format of columns to make them same across all datasets. (E.g.: Date and time format)
- Detect and remove duplicates (if any) and outliers (if any) in the datasets.
- Store a copy of the original and transformed datasets in multiple cloud platforms and systems for backup in case of failure of an individual system.

Dataset Description

- Mini Dataset – 128 MB with 286K records with 225 unique customers
- Large Dataset – 12 GB with 26M records with 22k unique customers

Model Selection

- Lasso Regression (for feature selection)
  - As we have many features in the above datasets, we will initially perform a Lasso regression on the above datasets to identify what features have strong correlation with our output variable.
  - This will give us a set of predictors that we will use to train our final mode.
- Logistic Regression, Linear Regression, Random Forest, Gradient boosting tree classifier (GBT) and support vector machine (SVM) (for predictions)
  - We are going to do a Logistic and Linear as we believe that will give us a good visualization of the ride counts, how it got affected during the covid period and the return to normal graph and easy to interpret.
  - We plan on using Random Forest, GBT and SVM as it might give strange results depending on the input predictors.

Tools

- Programming Language - Python
- Applications/Framework - Jupyter Notebook and Apache Spark
- Development - GitHub
- Visualization Tool - Tableau

METRICS TO MEASURE ANALYSIS RESULTS

Here are a few metrics to check the output results.

- Variance between predictors and response is measured by MSE, RMSE and R2 metrics.
- Accuracy metric will be used to check how accurate our model is by f1 score and ROC.
- RSE, VIF, F-stats and P-value for selecting the best model.

PROJECT PLANS AND MILESTONES

Task	Expected End Date	Owner	Status
Project Discussion 1	10/26/2022	Jasleen, Kranti, Nandish, Taranpreet	Completed
Understanding Project Topics	10/28/2022	Jasleen, Kranti, Nandish, Taranpreet	Completed
Project Topic Finalization	11/01/2022	Jasleen, Kranti, Nandish, Taranpreet	Completed
Finding Dataset	11/02/2022	Jasleen, Kranti, Nandish, Taranpreet	Completed
Project Proposal Submission	11/03/2022	Jasleen (Voice)	Completed
Project Discussion 2	11/07/2022	Jasleen, Kranti, Nandish, Taranpreet	Completed
Creating the Project Flow	11/11/2022	Jasleen, Kranti, Nandish, Taranpreet	In Progress
Setting up Environment	11/15/2022	Jasleen, Kranti, Nandish, Taranpreet	In Progress
Project Starting Report	11/21/2022	Jasleen (Voice)	In Progress
Dataset Cleaning	11/17/2022	Jasleen, Kranti,	In Progress
Data Visualization	11/17/2022	Nandish, Taranpreet	In Progress
Model Implementation	11/25/2022	Jasleen, Kranti, Nandish, Taranpreet	Pending
Testing and Validation	11/28/2022	Jasleen, Kranti, Nandish, Taranpreet	Pending
Conclusion	12/04/2022	Jasleen, Nandish	Pending
Future Scope	12/04/2022	Kranti, Taranpreet	Pending
Final Project Report	12/07/2022	Jasleen (Voice)	Pending

REFERENCES

- Shoro, Abdul & Soomro, Tariq. (2015). Big Data Analysis: Apache Spark Perspective. Global Journal of Computer Science and Technology. 15.
- Khedikar, Kanchan A., Data Analytics for Business Using Tableau (April 27, 2021). Proceedings of the International Conference on Innovative Computing & Communication
- Salman Salloum, Ruslan Dautov, Xiaojun Chen1, Patrick Xiaogang Peng1, Joshua Zhexue Huang, Big data analytics on Apache Spark
- N Balaji, B H Karthik Pai, Bhaskar Bhat and Barmavatu Praveen, Data Visualization in Splunk and Tableau: A Case Study Demonstration