CSI SUMMER INTERNSHIP 2020



**Project Report**

On

**APP SUBSCRIPTION ANALYSIS**

**Author**

Ankush Sharma

(Mahatma Gandhi’s Mission College of Engineering, Nanded)



**Co-Mentor**

Dr. A. V. Krishna Prasad

Dr. D.V. Ramana

**Mentor**

Mr. Naveen K.

BUSINESS UNDERSTANDING

* **Problem statement:**

There is a app which has basic as well as premium feature. The free version of app has basic feature. If the user wants to use premium feature then they have to pay some amount to unlock it.

The data contains customer’s behaviour and our job is to find insights form it.

We have to develop a model which predicts whether a customer buys a premium subscription of an app or not.

* **Motivation and Benefits from project:**
  + **Dataset**: Dataset has numerical, categorical as well as date-time variables which made dataset challenging and interesting.
  + **Critical thinking**: In this project, I was able to think like a professional developing a solution from a situation.
  + **In-depth exposure of Machine Learning:** In order to solve this problem, we had to refer to external sources instead of relying on training which provided exposure to different logics.
  + **Benefit:**
    - Improve myself in the field of data science.
    - Buy predicting the customers who will not be enrolling to premium to premium feature we can make good offers to them.
    - This will help us to increase app’s economy and increase enrolment.
* **Purpose:**
  + This project is part of CSI internship.
  + In this project, I am trying to predict customer’s response - whether to buy the premium version of an app or not.

DATA MINING

* **Data Source:** The source of the data is provided in form of link.

<https://drive.google.com/uc?export=download&id=1B1W303V3cBexAwf5Bw0H13IqM4gYwz3J>

* **Data set:** Dataset has 50000 records and 12 columns/feature that describes nature of user.

TECHNOLOGY

* **Google Colab**:  Colab is a free Jupyter notebook environment that runs entirely in the cloud.
* **Libraries:**
* **numpy :-**

Numpy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays

* **matplotlib :-**

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

* **pandas :-**

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

* **seaborn :-**

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

* **sklearn :-**

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms.

* **mlxtend-transaction-encoder :-**Transaction encoder is an encoding technique for converting list to columns. It is present in mlxtend package.mlx stands for Machine Learning Extension.

EXPLORATORY DATA ANALYSIS

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

* **Descriptive statistics**
  + Number of columns: 12
  + Number of rows: 50,000
  + Missing values: 18926
  + Duplicate rows: 15
  + Variable types: int64 - 8, object – 4
* **Descriptive statistics (after feature engineering)**
  + Number of columns: 120
  + Number of rows: 50,000
  + Missing values: 0
  + Duplicate rows: 15
  + Variable type: int64 - 120
* **Data visualization**
  + Number of user logged in to the app across the month.
  + Number of users logged in app across a week.
  + Number of users logged in across a day.
  + Histogram of user’s age.
  + Count Plot (columns): mini-game, enrolled, used-premium-feature, liked
  + Bar graph of different columns with respective to days of week.
  + Bar graph of different columns with respective to age.
  + Correlation matrix.
  + Correlation with response variable.
* **Points noted after data visualization**
  + There was slight increase in app’s growth.
  + Irrespective of users who didn’t like the app and extra feature like mini-games they enrolled for premium subscriptions.
  + From the graphs, we can say that users were more active in afternoon and evening as compare to the morning.
  + The most active users are between the ages of 25 – 35.
  + Enrolled column is mostly dependent on number of screens and list of screens.
  + The age, hour, num-screens, mini-game and used-premium-features are related

FEATURE ENGINEERING

Feature engineering is the process of using domain knowledge to extract features from raw data via data mining techniques. These features can be used to improve the performance of machine learning algorithms. Feature engineering can be considered as applied machine learning itself.

**Feature engineering methods:**

* **Parsing:** parsing is used to convert object into time-delta variable type, which is used to deal with data time columns and for removing outliners from the dataset.
* **Transaction Encoding:** It is an encoding technique to deal with screen list column in the dataset. It maps screens to fields (different columns).
* **Slicing** was used in hour column to convert it into integer type.
* Combined columns with same names.

For ex: ‘loan’, ‘loan1’, ‘loan2’ were combined and named as ‘loan count’.

* **Difference column** was generated by subtracting first open with enrolled date. This column tells us about ***time of response*** of a user to subscribe to premium feature. It is one of the most important column that helps us to deal with outliners
* Converted all entries of enrolled to 0 if their correspondence time of response is greater than 40. Hence dealt with outliners.
* **Dropped unimportant columns:** first open, enrolled date, difference (after dealing with outliners), users.

**Data Pre-processing**

Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

* **Feature scaling:** Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization.

In this project, I have used ***Standard scalar*** for feature scaling.

* **Splitting of dataset into train and test:** In this dataset is divide into two parts – train and test.

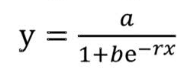
***Train subset*** is for building the model.

***Test subset*** is for predicting and finding the accuracy.

In this project, I have used train-test-split of pre-processing of sklearn package.

PREDICTIVE MODLELING

* Logistic Regression:
  + - Regression is a method to determine a statistical relationship between dependent variable and one or more independent variables.
    - Logistic regression is a statistical model on which the response variable takes a discrete value and the explanatory variable can be either continuous or discrete.
    - Logistic Regression model is given by



* I have used logistic regression because it is an appropriate regression analysis to conduct when dependent variable is dichotomous (binary).
* Decision Tree Classification:
  + Decision Tree Learning are a collection of divide and conquer problem-solving strategies that use tree-like structure to predict the value of an outcome variable.
  + The tree starts with a root node consisting of complete data and thereafter uses intelligent strategies to split the nodes into multiple branches.
  + This is done to create more homogeneous groups at the children nodes.
  + *Gini**index*and *Entropy* are the criterion of decision tree model for calculating information gain.
  + Both are measures of impurity of a node.
* **Grid search cross validation**

***Grid-search*** is process of scanning and is used to find the optimal hyper-parameters of a model which results in the most ‘accurate’ predictions.

RESULT

* Logistic Regression accuracy: **87.15 %**
* Decision Tree Classification accuracy: **87.22 %**
* Decision Tree Classification accuracy (after hyper-parameter tuning): **87.13 %**

CONCLUSION

Logistic Regression and Decision tree classifier are two most popular and basic classification algorithm being used today.

But none of the algorithm is better than the other and one’s performance is often credited to the nature of data being work upon.

Thus, we the help of this model, we will be able to predict the users who are not willing to enrolled for premium subscription.

We can increase efficiency by providing them with good offers.

REFERENCE

<https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc>

<https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4>