ABSTRACT

The aim of the project to investigate and compare the performance of different machine learning (ML) methods as applied to three different data sets. The datasets originate from Google Projects and services. The Project sunroof is an initiative to calculate solar energy potential to via aerial imagery. Regression model was applied to predict solar potential for all roof types in a region. The random forest regressor algorithm gave the R Squared value of 0.98 for the same. Play store is the certified application store for android certified mobile devices. Regression model was implemented to the dataset for predicting the characteristics of an application. The random forest regressor algorithm gave the R squared value of 0.89 when predicting the rating for the application. YouTube is online video platform with millions of stream able videos. Classification ML model was applied to predict the category of video on a YouTube channel. XGB ML model gave the accuracy score of .95 when predicting the category for a video on T-Series YouTube channel. It was found that random forest regressor was the optimum ML method for regression while XGB classifier method was the better model for classification. The author is most familiar with coding language Python; hence it is used primarily in the project.

INTRODUCTION

The implementation of machine learning (ML) has seen an exponential increase in all fields of work as well as services. This is due to the inbuilt ability of the ML to elucidate answers which thereby are difficult to learn otherwise. The report in general aims to use this inherent feature of ML to answer interesting question by using regressing and classification models on interesting topics. The topic of choice for this report is the Google services and projects. Inline, Google Sunroof Project [1], Google Play Store Apps [2] and T-Series video channel on Goggle YouTube were chosen as datasets. The aim is to develop multiple ML models for each case to predict the target variable which gives the most efficiency, precision, and accuracy.

Project sunroof

While starting as a part-time project initiative of Google engineer Carl Elkin, the Project Sunroof has become valuable source of information regarding solar energy. The initiative employs aerial and satellite photography to map a region and provide potential solar panel location on roofs and land. The data from Google Maps and Google Earth is used to calculate shadows and dark areas from nearby infrastructure and landscapes to find the optimal location to set up solar panels. The project sunroof dataset was chosen because of the large number of rows and columns as well as the organised records. The Research question for this dataset is as follows:

Play Store

In 2019, Google android operating system accounted for 87% of all certified mobile globally. Google Play store being the official pre-installed application store for android is the most used android app in the world. It also holds 2.65 Million downloadable application. The google dataset was chosen because of the immense records and information. The research question for this dataset is given below:

User Ratings is the biggest feedback factor which is used by developers to evaluate their applications. Hence predicting the User rating is very important in terms of ad revenue which is usually the main source of income in app development. Successfully predicting the rating will also give insight on the most corelated parameter that makes an app successful.

YouTube

Headquartered in San Bruno, YouTube is the largest video streaming service in the world. With million of videos available online and millions of people watching regularly, Youtube has been used small and large production companies to grow audiences. For this project, T-Series channel dataset was used to implement ML model. T-Series is a Music Production Company in India with the highest subscribers count on YouTube globally. The research questions are: