Machine Learning Analysis on various Google Entities

Aditya Kumar Singh-x20140410
MSc Data Analytics - Jan 2021, Group B
National College of Ireland
Dublin, Ireland

***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.**

***Index Terms*—Machine learning, data sets, data sources, evaluation methods, regression, classification**

I. Introduction

The implementation of machine learning (ML) has seen a 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 [3] 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.

A. Google Sunroof Project

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 organised records. The Research question for this dataset is as follows:

1) Can we implement a model to predict the kW of solar potential for all roof types in a region?

It is important to to predict potential benefits before any investment is placed. Hence considering the immense initial costs setting up solar panels means that enough study needs to be conducted before any infrastructure is placed. Predicting the total energy potential that can be captured is of utmost important for a feasible project.

B. Google 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:

1) Can we use a model to predict the user rating for a app among the record?

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 co-related parameter that makes an app successful.

C. T-Series Google YouTube Channel

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:

1) Can we implement a model to predict the video category of a video on channel?

For channels on YouTube, the source of revenue stream from the number of audiences who actively watch their videos. To find the most successful genres of video it is important to classify the best category of videos.

II. Related Works

This section will discuss the related research papers and other works that were considered while working on the stated research questions.

As mentioned earlier, google play is one of the most used android applications in the world. The [4] is an in-depth research study on the reviews scrapped from the app reviews of the google play store. The authors implemented ML models to classify the semantics of the text as negative, positive, or neutral. The study was able to verify Logistic regression as the better model based on F1 score and accuracy. The limitation of the paper is one dimension focus on text reviews and feedback while not considering the fact that most users prefer to leave start rating. The Google Play Store gives the point-based feedback system. The comments are often given by users with negative experience. This makes study biased.

While the previous paper focused on text reading, [5] is research paper with direct study topic with paper. The study is conducted to predict movie user ratings using different ML methods. The study showcased that Latent Variable Model (LVM) is the better model for predicting audience ratings for movies. While the paper is successful on implementing ML methods that can predict ratings, the usual database used is limited to one region and lacks diversity to show the large scale implementation.

The [6] discusses one of the benefits of user ratings which is inferring the demography of the audience of the user. This is a useful feature to provide directed services and products. The paper discusses the potential new model for this specific type of ML prediction.

The [7] elucidates about the discrepancy between numeric rating and reviews. The paper applies ML model on the average semantics to portray the numeric ratings and then compare them with the source numeric ratings. The classification model used here is random forest classifier. The study shows that User Ratings are often manipulated and do not coincide the true user experience. This is a useful topic for further research since the project aims to predict the user rating on Google apps. It is thus necessary to make sure that the ratings are genuine and not manipulated by anyone.

The [8] focuses on the user behavioral pattern for the user ratings. The study suggests that some users are more inclined to give ratings as compared to others and hence it is important to normalize for unbiased study. The paper uses Gaussian Normalization and Collaborative Filtering method to reach conclusion on the research question. The paper concludes with Collaborative Filtering more accurate as compared to other.

Solar energy has had deep research for the past few decades. The amount of related works on solar energy is plentiful. As such,[9] discusses the prediction solar power generation from weather forecast. This is in-line with our study that also use the average sunlight received as a predictor variable.

The [10] is a research study that used various ML models and compare the results in order to find the best predicting model to Solar Radiation. The study concludes with SVM as the better model for the research.

While most studies prefer traditional models for solar energy prediction, [11] uses Extreme Learning Machine (ELM) to map daily global solar radiation. The paper studies photovoltaic power quantification. The lack of ELM methods in python proved to be an hindrance in this project. The inability to compare random forest model with ELM needs to be done at a later stage.

Machine learning related works in Solar predictions often compare only among ML models. [12] includes Artificial Neural Networks (ANN) for comparative study with ML models namely SVM. This can be done be future evaluation of this project. The study shows the out-performance of ANN with respect to ML models.

The project uses roof type, number of panels on roof for predicting the potential solar energy generation. This is also discussed in [13], where the classification ML model is used for classifying roof shapes in Geneva for solar energy applications. The studies agree with flat roof being the most efficient for solar energy applications as compared to other roof types.

The paper[14], I the basis of choosing XGBoost as a ML model for project. Predicting the popularity is essential for any revenue generating company which deals with content manufacturing. This also shows the most popular video category.

From a content creator perspective[29], learning the key to audience taste the most profitable aspect that they can predict. [15] discusses the decision support system that can be employed for online video channel. The study helps in topic consideration for the project.

An important aspect of machine learning includes the prediction of human emotions. The paper [16] classifies the videos on emotion factors. The study focuses on the ability of machine learning to predict complex prediction variables. This paper was useful for topic consideration for the project.

The [28] Project Sunroof relays great importance on the shadows cast my land features on roofs where solar panels can be constructed. [17] improves upon the machine learning aspects of Google Earth data. The project relies on the dataset from Google Earth, hence a study inspection to get a view of ML models was necessary.

City infrastructure shape sunlight for low laying houses and lands. This has implication for any solar panels that are to be constructed there. [18] investigates a ML model for urban to model solar irradiation.[26]

The study [19] has been a significance influence on the project for choosing XGBoost as a classifier for YouTube Dataset. The study shows the superior performance of the XGBoost as compared to other models. This is inline with the project results.

Random forest decision tree has the highest performance parameters in the project study. This is in-line with [20] conclusion. This is critical for the model selection for the project. The study works upon the efforts for improving the current model. There is a comparison with logistical regression.

The project focuses much on the data generated because of the online activity. [21] elucidates on the role of demographics in online learning. The paper shares light on the decision tree machine learning model. The focus is much on the effect the different demography shed on the online activity.

The [22] on Kaggle discusses the same topic as the project. The work here is limited to linear regression. The author concludes that the there is not enough correlation between predictors variable to apply linear regression for a successful model. [25] The project wants to further this study and apply other models to see if the previous works can be confirmed and added upon.

III. Methodology

A. Google Sunroof Project

1) Dataset: The unprocessed dataset contains 10,622 Rows and 31 columns. It contains information regarding the solar panels and roofs with their potential solar energy.

2) Methodology: The applied methodology for preprocessing the sunroof dataset was Knowledge Discovery in Databases (KDD). This is in-line with research done for the same. The models for used in this dataset were linear regression, random forest regressor and decision tree regressor.

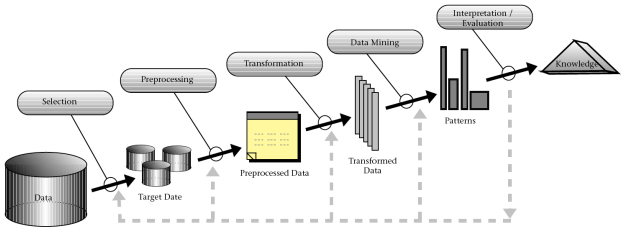

Fig. 1. KDD Methodoly Flow Chart

1) Data Selection and Cleaning: The Dataset project sunroof was downloaded from kaggle. The dataset was checked for missing or null values. All rows that showed otherwise were dropped.
2) Data Preprocessing and Transformation: Irrelevant and unnecessary features such as 'installsizekwbucketsjson' were dropped from data. 'statename' column was transformed into numeric values using label encoder. Pearson Correlation Matrix graph was plotted to find the the correlated rows with the target variable.

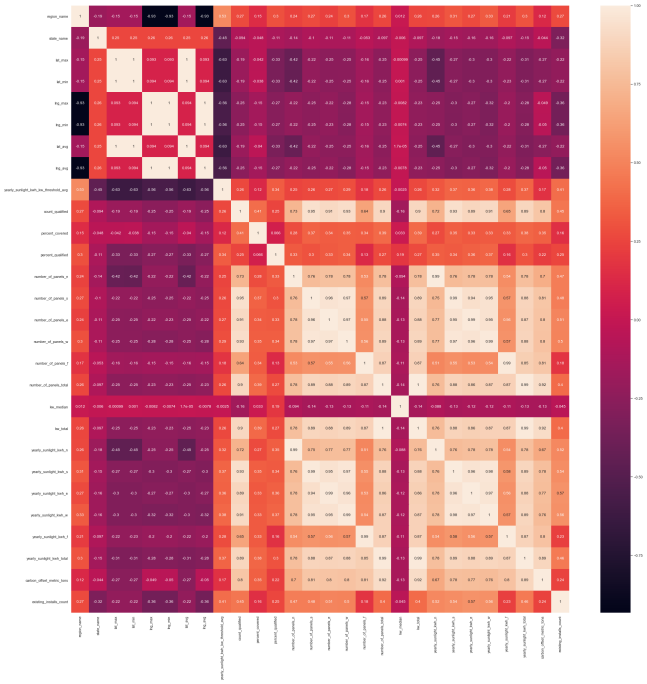

Fig. 2. Correlation Matrix for Project Sunroof Dataset

As seen from the plot in Figure 2, there are various columns that showed extremely high correlation with the target variable. The columns that showed collinearity of 1 with target variable were found to be 'numberofpanelsmedian' and 'yearlysunlightkwhmedian'. These columns were dropped from the datasets as it would not be impossibe to implement a model. Several other columns which showed multicollinearity can also be seen in Figure 1. 'numberofpanelsf','yearlysunlightkwhf' and 'yearlysunlightkwhtotal' were found to be the columns that show high collinearity with the target variable. These were dropped for as to not obtain an overfitting model. Columns that showed multicollinearity with each other were also found and dropped. As seen in Figure 2, after dropping all the unnecessary variable columns, the new matrix ready for model implemetation.

3) Model Fitting: The data was divided into test and train for models to fit. For this new dataframes X and y were created where X contains predictors and y contains target variables. It is necessary to implement so as to get parameters for each regression model.
The dataset was split on the ratio of 80:20 where 80 is train and 20 is test.
After splitting, linear regression model and random forest regressor model were fitted and their result evaluated.

B. Google Play Store

1) Dataset: The unprocessed dataset contains 1,11,393 Rows and 23 columns. The dataset was ontained online from kaggle website. The original dataset contained 1.1 million

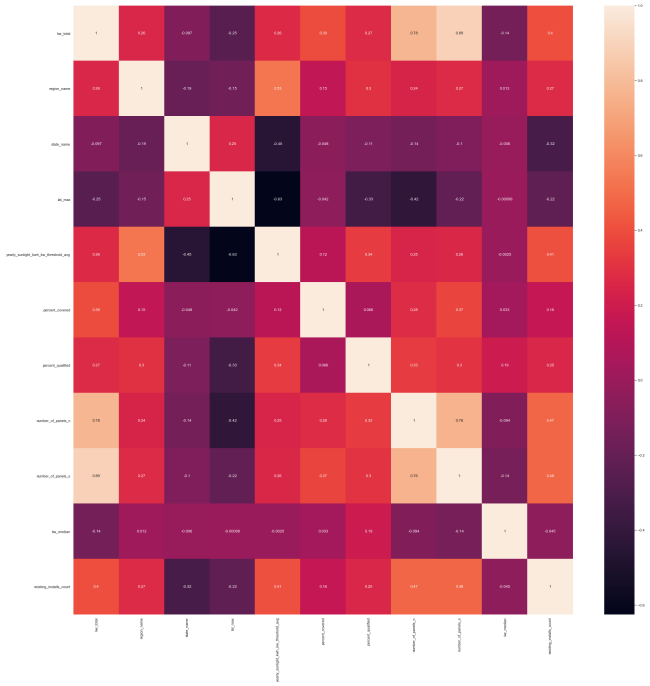

Fig. 3. Correlation Matrix for Project Sunroof Dataset after dealing multi-collinearity

```
# split into train test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

```
(8468, 11) (2117, 11) (8468, 1) (2117, 1)
```

Fig. 4. Splitting Dataframes

records. For the project a smiliar but smaller dataset was chosen.

2) Methodology: The applied methodology for preprocessing the sunroof dataset was Knowledge Discovery in-Databases (KDD). This is in-line with research done forthe same. The models for used in this dataset were linearregression, random forest regressor and decision tree regressor.

1) Dataset Selection and Cleaning: After initial cleaning of null vaues and duplicates, the new dimension retrieved 80891 rows.
2) Preprocessing and Transformation : Then unneccessary columns such as App Id,Developer Email,Developer Website,Privacy Policy and Size were dropped. These columns had no relevance to prediction the purpose in the project. Catogorical columns were transformed via label encoder.
   After transformation correlation matrix was plotted to check multicollinearity.
   As seen in the plot in Figure 6, there is only 2 columns that show multicollinearity.Hence one of them was dropped. All the other columns can be used to perform modeling. The dataset was split into separate dataframes as same like the previous datasets. Then the dataframs were further split into test and train for evaluation. For this dataset, we performed Linear Regression, Random

```
XY = data_no_dupl[['kw_total','region_name','state_name','lat_max','yearly_sunlight_kwh_kw_threshold_avg','percent_covered','perc
X  =  data_no_dupl[['region_name','state_name','lat_max','yearly_sunlight_kwh_kw_threshold_avg','percent_covered','percent_qualif
y = data_no_dupl[['kw_total']]
```

Fig. 5. Creating Dataframe for Test and Train

```
from sklearn.linear_model import LinearRegression
lin = LinearRegression()
lin.fit(X_train, y_train)
ylin_pred = lin.predict(X_test)
from sklearn import metrics

print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, ylin_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, ylin_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, ylin_pred)))
print('R2 score:',metrics.r2_score(y_test, ylin_pred))
```

```
Mean Absolute Error: 1.1927151296695764
Mean Squared Error: 2.41300967785878
Root Mean Squared Error: 1.5533865191441505
R2 score: 0.05798500416871433
```

Fig. 6. Linear Regression Code Used

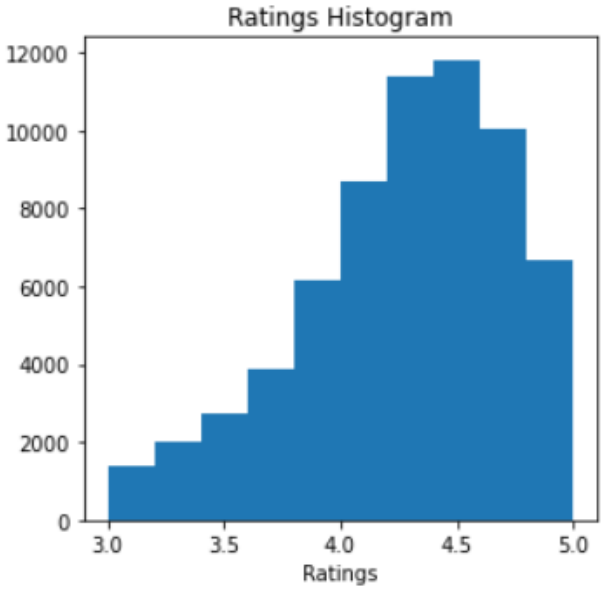

Fig. 7. Rating Distribution

```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

data_no_dupl['Free']= le.fit_transform(data_no_dupl['Free'])
data_no_dupl['Developer Id']= le.fit_transform(data_no_dupl['Developer Id'])
data_no_dupl['App Name']= le.fit_transform(data_no_dupl['App Name'])
data_no_dupl['Category']= le.fit_transform(data_no_dupl['Category'])
data_no_dupl['Installs']= le.fit_transform(data_no_dupl['Installs'])
data_no_dupl['Content Rating']= le.fit_transform(data_no_dupl['Content Rating'])
data_no_dupl['Ad Supported']= le.fit_transform(data_no_dupl['Ad Supported'])
data_no_dupl['In App Purchases']= le.fit_transform(data_no_dupl['In App Purchases'])
data_no_dupl['Minimum Android']= le.fit_transform(data_no_dupl['Minimum Android'])
data_no_dupl['Editors Choice']= le.fit_transform(data_no_dupl['Editors Choice'])
data_no_dupl['Currency']= le.fit_transform(data_no_dupl['Currency'])
```

Fig. 8. Encoding Categorical Column into numerical values

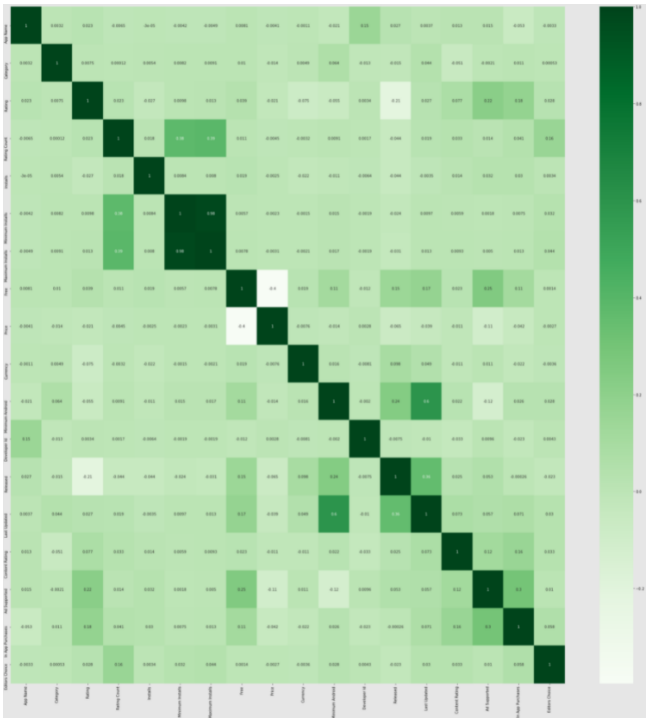

Fig. 9. Correlation Matrix for Google Playstore Dataset

Forest Regressor and Decision tree regressor.

Random Forest Regressor

```
from sklearn.ensemble import RandomForestRegressor
X =  solar[['region_name','state_name','lat_max','yearly_sunlight_kwh_kw_threshold_avg','percent_covered','percent_qualified','r
y = solar[['kw_total']]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=1)
classifier = RandomForestRegressor(n_estimators = 10)
classifier.fit(X_train, y_train.values.ravel())
yRand_pred = classifier.predict(X_test)

print('Mean Absolute Error:', metrics.mean_absolute_error(y_test,yRand_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test,yRand_pred))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test,yRand_pred)))
print('R2 score:',metrics.r2_score(y_test, yRand_pred))
```

```
Mean Absolute Error: 20227.814182805858
Mean Squared Error: 1131509472.1024265
Root Mean Squared Error: 33637.91717842272
R2 score: 0.8088936824254577
```

Fig. 10. Random Forest Regressor Code Used

C. Google YouTube T-Series

1) Dataset: The data set was downloaded from kaggle website. The dataset is a comprehensive collection of videos on T-Series youtube channel. The dataset contains video description of videos uploaded by the company with their view ounts, likes , ratings and dislikes. The records available are in 14,302 columns and 14 rows.

2) Methadology: Just like the others, KDD model was implemented for this dataset. The model that are to be used are Random Forest Classifier and XGBoost.

1) Preprocessing and Transformation: The dataset was cleaned and null and duplicat values dropped. The resulting dataset consisted of 14,239 rows. Columns that had no relevance to machine learning were dropped. In this datae t 'videoCategoryId' column was dropped. Categorical columns like 'videocategorylabel', 'definition' and 'caption' encoded to numerically. after this multicollinearity check was done by plotting the correlation matrix. As observed in the plot in Figure 8, there is

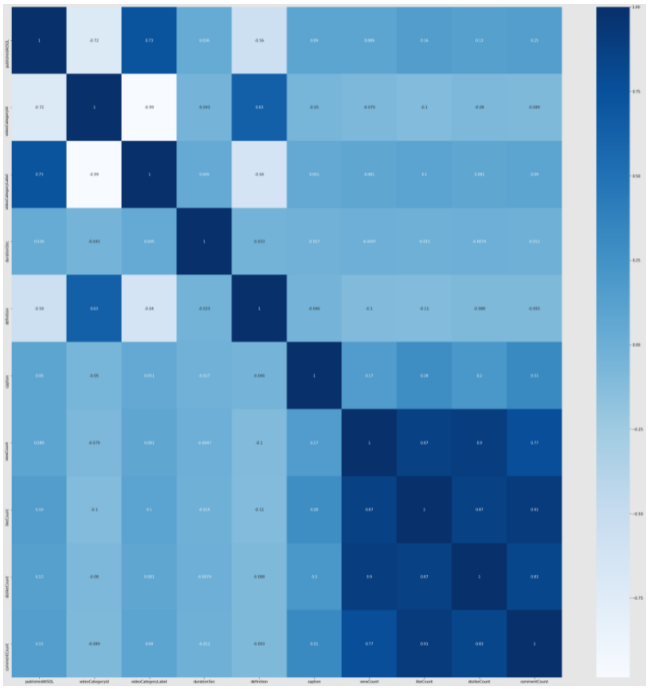

Fig. 11. Correlation Matrix for Yotube Dataset

high collinearity between target and predicting variable. There is also multicollinearity between some predciting variables as well. These were dropped as a result for a perfect model.

2) Model Fitting: The dataset was again divided into dataframes. The dataframes were then again splitted in to Test and Train dataframes. as compared to regression models, this time the spilt was in the rtion 70% and 30%. For classification Decision Tree Classifier, Random Forest Classifier and XGBosst models were used.

Decision Tree

```
from sklearn import metrics
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier

clf = DecisionTreeClassifier()
# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)
#Predict the response for test dataset
y_pred = clf.predict(X_test)
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
from sklearn.metrics import f1_score
print("F1 Score:",f1_score(y_test, y_pred, average='macro') )
```

```
Accuracy: 0.9375
F1 Score: 0.6603035617368519
```

Fig. 12. Decision Tree Classifier Code Used

IV. Evaluation

Here we will evaluate the results different model that were implemented. For regression we used Linear Regression, Random Forest Regressor and Decision Tree Regressor.The metric that we used for comparison are Mean Absolute Error, Mean Squared Error,[23] Root Mean Squared Error and R2 score. This is the final part of KDD where interpretation of result is discussed and the significance of values obtained.

A. Google Project Sunroof Result Evaluation

Below in the Figure 9 are the results obtained from the all the models. As seen here Random Forest Regressor is the best model for prediction with R- squared value of 0.81 and the lowest Mean Absolute Error. In comparison the Decision Tree Regressor performed the worst with highest Mean Absolute Error and R squared score of 0.64. Linear Regression also performed worse as compared too Random Forest Regressor.

Decision Tree Regressor

```
Mean Absolute Error: 26730.636041568257
Mean Squared Error: 2108739823.6283066
Root Mean Squared Error: 45921.017232072576
R2 score: 0.6438443403680874
```

Random Forest Regressor

```
Mean Absolute Error: 19768.52083136514
Mean Squared Error: 1085420445.5222878
Root Mean Squared Error: 32945.719684388256
R2 score: 0.816677889599578
```

Linear Regression

```
Mean Absolute Error: 20871.892505097345
Mean Squared Error: 1116549967.0619335
Root Mean Squared Error: 33414.81657980384
R2 score: 0.8114202683635442
```

Fig. 13. Different Metrics for Google Project Sunroof Dataset

The following result shows the inability of linear regression to perform efficiently when high numbers of predictors are used. This also shows that Decision tree and Random forest are able to handle higher number of predictors variable well as compared to Linear Regression[24].

B. Google Play Store

Below in the Figure 10 are the results obtained from the all the models for Google Play store data. As seen here Random Forest Regressor is the best model for prediction with R-squared value of 0.89 and the lowest Mean Absolute Error. In comparison the Decision Tree Regressor performed the worse as compared to Random Forest Regressor with highest Mean Absolute Error and R squared score of 0.81.

One interesting aspect is the R squared score of Linear Regression in the results. The score of 0.05 is inline with [22] where the author also recieved the R squared score 0.4. The decision tree and Random forest were able to give results because of the nature of their algorithm.

Decision Tree Regressor

```
Mean Absolute Error: 0.4448297175350764
Mean Squared Error: 0.46883552753569435
Root Mean Squared Error: 0.6847156545133858
R2 score: 0.8169712697095353
```

Random Forest Regressor

```
Mean Absolute Error: 0.34227455343346314
Mean Squared Error: 0.263829260152049
Root Mean Squared Error: 0.5136431252845199
R2 score: 0.8970036789811628
```

Linear Regression

```
Mean Absolute Error: 1.1927151296695764
Mean Squared Error: 2.41300967785878
Root Mean Squared Error: 1.5533865191441505
R2 score: 0.05798500416871433
```

Fig. 14. Different Metrics for Google Play Store Dataset

C. T-Series Google YouTube Channel

For T-Series Google Youtube Dataset the result are given in Figure 11. As can be seen, both the Random forest Classifier and Decision tree gave the accuracy of 0.93. In comparison the XGBoost have the highest accuracy of 0.94. When we consider the F1 score Decision Tree has the highest score with 0.66. XGBoost has has slight down score of 0.62 while the Random Forest Classifier has the lowes score of 0.61.

For selecting the best model, XGBoost can be considered for more effeciency and fast performance. It is also the most accurate. F1 score is the geometric mean of precision and recall. Hence while Decision tree classifier gives higher precision, the result may not be accurate.

V. Conclusion and Future Works

The goals of the project as mentioned in research have been met. The author hopes that unlike various Machine Learning projects which are based on common topics, exotic datasets are worked upon. The topic of this project hopefully is rare with very less research and related work available.

The prediction of potential solar energy is an important research topic and hopefully more research will be done in the future. More complex methods will be used for evaluating the this interesting subject as renewable resource for energy are coming positive light.

The subject of Ratings and Reviews is an interesting behavioural subject in its own. The customer approach and reviews to a specific product plays an important role in deciding the developer's approach towards a specific presentation.

More efficient and powerful machine learning methods can be impelmented in future on the same subjects.

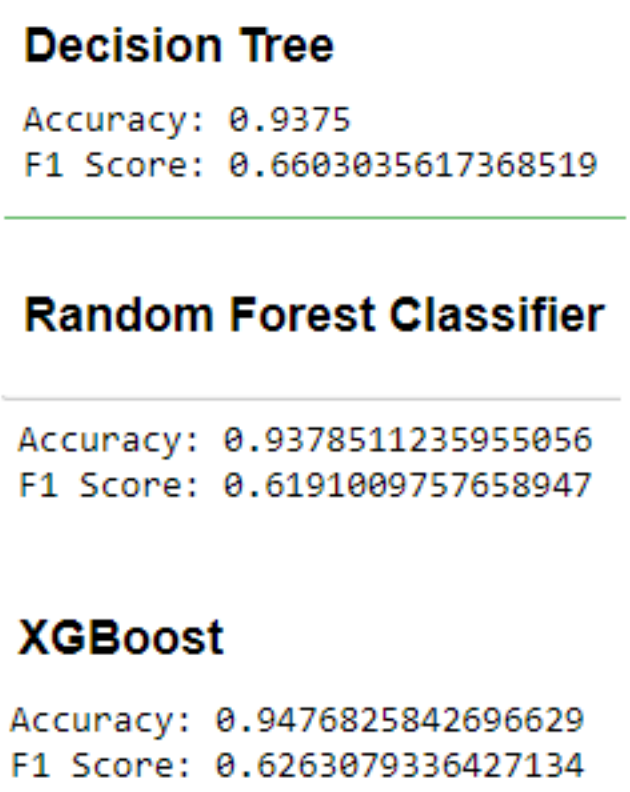

Fig. 15. Different Metrics for T-Series Google YouTube Channel Dataset

REFERENCES

[1] Google Project Sunroof, [Online], Available: https://www.kaggle.com/jboysen/google-project-sunroof

[2] Google Play Store Apps, [Online], Available: https://www.kaggle.com/gauthamp10/google-playstore-apps

[3] PewDiePie vs T-Series, [Online], Available: https://www.kaggle.com/phiitm/pewdiepie-vs-tseries

[4] Abdul Karim and Azhari Azhari , "Methodology for Analyzing the Traditional Algorithms Performance of User Reviews Using Machine Learning Techniques," Advanced Data Mining: Algorithms and Applications, vol. 27, pp. 141–157, August 2013.

[5] Mladen Marovic, Marko Mihokovi ´ c, Mladen Mik ´ sa, Sini ˇ sa Pribil, and Alan Tus,"Computers in Human Behavior, Volume 59, June 2016, Pages 165-172.

[6] Peter Brauna, Alfredo Cuzzocreab, Lam M.V. Doana, Suyoung Kima, Carson K. Leunga, Jose Francisco A. Matundana and Rashpal Robby Singh, "Enhanced prediction of user-preferred YouTube videos based on cleaned viewing pattern history," International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France.

[7] Joon Soo Lim, Min-Ji Choe, Jun Zhang and Ghee-Young Noh, "The role of wishful identification, emotional engagement, and parasocial relationships in repeated viewing of live-streaming games: A social cognitive theory perspective ", Computers in Human Behavior, Volume 108, July 2020, 106327.

[8] Muhannad Quwaider, Abdullah Alabed and Rehab Duwairi, "The impact of video games on the players behaviors a survey ", The 10th International Conference on Ambient Systems, Networks and Technologies (ANT), April 29 – May 2, 2019, Leuven, Belgium.

[9] Video Game Sales 2020, [Online], Available:

[10] Yu Feng, Weiping Hao, Haoru Li, Ningbo Cui, Daozhi Gong, Lili Gao, "Machine learning models to quantify and map daily global solar radiation and photovoltaic power", Renewable and Sustainable Energy Reviews, Volume 118, 2020, 109393, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2019.109393.

[11] Imane Jebli, Fatima-Zahra Belouadha, Mohammed Issam Kabbaj, Amine Tilioua,"Prediction of solar energy guided by pearson correlation using machine learning", Energy, Volume 224, 2021, 120109, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2021.120109.

[12] Nahid Mohajeri, Dan Assouline, Berenice Guiboud, Andreas Bill, Agust Gudmundsson, Jean-Louis Scartezzini, "A city-scale roof shape classification using machine learning for solar energy applications",Renewable Energy, Volume 121, 2018, Pages 81-93, ISSN 0960-1481, https://doi.org/10.1016/j.renene.2017.12.096.

[13] Hongjun Jeon, Wonchul Seo, Eunjeong Park, Sungchul Choi,"Hybrid machine learning approach for popularity prediction of newly released contents of online video streaming services", Technological Forecasting and Social Change, Volume 161,2020, 120303, ISSN 0040-1625, https://doi.org/10.1016/j.techfore.2020.120303.

[14] Stephen L. France, Yuying Shi, Mahyar Sharif Vaghefi, Huimin Zhao, "Online video channel management: An integrative decision support system framework", International Journal of Information Management, 2020, 102244, ISSN 0268-4012,https://doi.org/10.1016/j.ijinfomgt.2020.102244.

[15] Yen-Liang Chen, Chia-Ling Chang, Chin-Sheng Yeh, "Emotion classification of YouTube videos",Decision Support Systems, Volume 101, 2017, Pages 40-50, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2017.05.014.

[16] Lamuel Chi Hay Chung, Jing Xie, Chao Ren, "Improved machine-learning mapping of local climate zones in metropolitan areas using composite earth observation data in Google Earth Engine", Building and Environment, 2021, 107879,ISSN 0360-1323, https://doi.org/10.1016/j.buildenv.2021.107879.

[17] Vartholomaios, Aristotelis," A machine learning approach to modelling solar irradiation of urban and terrain 3D models", Computers Environment and Urban Systems, 78,101387101016

[18] Jitendra Shreemali, Lokesh Malviya, Payal Paliwal, Prasun Chakrabarti, Sandeep Poddar, Bhavesh Jindal, Harshit Chaubisa, "Comparing performance of multiple classifiers for regression and classification machine learning problems using structured datasets", Materials Today: Proceedings, 2021, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2020.12.480.

[19] R. Madana Mohana, C. Kishor Kumar Reddy, P.R. Anisha, B.V. Ramana Murthy,"Random forest algorithms for the classification of tree-based ensemble, Materials Today: Proceedings",2021,ISSN 2214-7853,https://doi.org/10.1016/j.matpr.2021.01.788.

[20] Saman Rizvi, Bart Rienties, Shakeel Ahmed Khoja,"The role of demographics in online learning; A decision tree based approach",Computers Education, Volume 137, 2019, Pages 32-47, ISSN 0360-1315,https://doi.org/10.1016/j.compedu.2019.04.001.

[21] A. Moosa, H. Shabir, H. Ali, R. Darwade and B. Gite, "Predicting Solar Radiation Using Machine Learning Techniques," 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, pp. 1693-1699, doi: 10.1109/ICCONS.2018.8663110.

[22] N. Sharma, P. Sharma, D. Irwin and P. Shenoy, "Predicting solar generation from weather forecasts using machine learning," 2011 IEEE International Conference on Smart Grid Communications (SmartGridComm), 2011, pp. 528-533, doi: 10.1109/SmartGridComm.2011.6102379.

[23] J. S. Breese, D. Heckerman and C. Kadie, Empirical Analysis of Predictive Algorithms for Collaborative Filtering, Proc. of the 14th Conference on Uncertainty in Artificial Intelligence (UAI), 1998

[24] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl, Grouplens: An Open Architecture for Collaborative Filtering of Netnews. In Proc. of the ACM 1994 Conference on Computer Supported Cooperative Work, pages 175–186

[25] T. Hofmann, Gaussian Latent Semantic Models for Collaborative Filtering, In the Proc. of the 26th Annual International ACM SIGIR Conference, 2003

[26] L. Si and R. Jin, Flexible Mixture Model for Collaborative Filtering, In Proc. Of the 20 th International Conference on Machine Learning (ICML 2003), 2003

[27] R. Jin, L. Si, and C.X. Zhai, Collaborative Filtering with Decoupled Models for Preferences and Ratings, in Proc. Of the 12th Conference on Information and Knowledge Management (CIKM 2003), 2003.

[28] M. Marović, M. Mihoković, M. Mikša, S. Pribil and A. Tus, "Automatic movie ratings prediction using machine learning," 2011 Proceedings of the 34th International Convention MIPRO, 2011, pp. 1640-1645.

[29] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions", IEEE Trans. Knowl. Data Eng., pp. 734-749, 2005.