***GOOGLE PLAYSTORE APP RATING USING VARIOUS ML ALGORITHMS***

**Introduction**

I am extremely excited to explain and share my thought process on how to Predict Google Play Store App Ratings. We still start off with basic Data Mining definitions and get into algorithms with our Google Play Store dataset.

**Data Mining**

Data mining is the process of looking through a data set and finding correlations, anomalies and or patterns that can be of usefulness. In other words, it is having a large dataset full of scattered information and trying to make sense of it by finding meaningfulness. For example, if a company like Macys wanted to predict the sales of a new item, they could look at similar items.

**Python**

Most of the data scientist use python because of the great built-in library functions and the decent community. Python now has 70,000 libraries. Python is simplest programming language to pick up compared to other language. That is the main reason data scientists use python more often, for machine learning and data mining data analyst want to use some language which is easy to use. That is one of the main reasons to use python. Specifically, for data scientist the most popular data built in open source library is called panda. As I have seen earlier in my previous assignment when I need to plot scatterplot, heat maps, graphs, 3-dimensional data python built-in library comes very helpful.

**Software**

Use Anaconda Navigator to launch Jupyter Notebook. Then we choose Python 3 for coding.

**Data**

There are various ways of finding the data. Most ideal way to find data is to search it online. There is good amount of website that can provide large data, for example, Data.Gov, The Federal Reserve, Kaggle.

My Dataset

Kaggle: https://www.kaggle.com/gauthamp10/google-playstore-apps?select=Google-Playstore-Full.csv

Dataset was obtained, by scrapping off Google Play Store.

**Domain Knowledge**

It is hard to do projects in a domain you do not know or understand. You should understand the situation of the data and the importance of it. Suppose you want to build a model to predict how frequent oil changes are need Japanese cars. You should know the domain to understand the variables to gather, to use the data to predict the outcome. If you understand the domain you will understand false positives, outliers, and better measure errors. If you do not have a domain about car manufacturing, parts, vehicle models, etc. you will not be able to make a model. After you get the data and model, you should understand and analyze the data so you can communicate your findings. These are just a few important reasons why domain knowledge is important.

**Body**

Now you have introduction of Data Mining, let’s get into it!

**Hypothesis**

Predict Google Play Store app ratings.

***Independent Variables***

Independent variables that do not have any dependence variables that can change the outcome of independent variables. Another name for independent variables is Predictor variables. In data, independent variables will also be known as regressors, controlled variables, manipulated variables, explanatory variable, exposure variable, and/or input variable. Most of the time the companies want to control independent variable to see the relationship of the dependent variables factors and the outcome.

**Dependent Variables**

Dependent variables are the opposite. There outcome is dependent on another variable. Dependent variables can measure the effects of the independent variables. Dependent variables usually depend on the independent variables. Another name for dependent variables is Predicted variables. In data, dependent variables will also be known as response variable, regressand, measured variable, observed variable, responding variable, explained variable, outcome variable, experimental variable, and/or output variables.

**Preprocessing**

Preprocessing is important into transitioning raw data into a more desirable format. Undergoing the preprocessing process can help with completeness and compellability. For instance, you can see if certain values were recorded or not. Also, you can see how trustable the data is. It could also help with finding how consistent the values are. We need preprocessing because most real-world data are dirty. Data can be noisy i.e. the data can contain outliers or just errors in general. Data can also be incomplete i.e. there can be some missing values.

Data storage is getting easier and easier, most organizations are storing a bulk of data, for various purposes like auditing, information, prediction etc. Which turns out to be beneficial to the organization. But at the same time, bulk data said as big data increases the possibility of noisy data. Also, as this data would be coming from different sources, there is a higher chance of inconsistency. Even in smaller data set there are missing entries and inconsistent formats. All these anomalies hamper the cause of storing data, it can cause low quality mining, the prediction analysis might get drifted to the wrong side. While violets the cause of storing data, hence it is important to make the data as required which we say as preprocessing. This brings in data accuracy, completeness, consistency, timeliness, believability and interpretability.

**Algorithms**

There are many algorithms we can use on our dataset. Some algorithms will work better for you over others. You will just have to try few different algorithms for your dataset, and see which one gives you the best results. Here are few popular algorithms we used that worked best for our dataset.

**Random Foresting**

Random forest is basically creating a forest of decision trees and make it random. We will see a relationship between our trees and the result we are going to get. For example- If I am deciding what movie to watch this Sunday, and I want the suggestion from my friends. My friends will ask me random question like what genre first and then what specific type of movies from that genre. So, over here friend’s suggestions are treated as decision tree. Finally, I am going to choose a movie who got the most votes. That is the result. These a random foresting algorithm example.

Some of the advantages and disadvantages of random forest classifiers are as follows:

**Advantages**

One advantage of random forest model is it never over-fits the model. It can handle missing values. We can use Random forest for both classification and regression. The random forest can help us identify the most important part from our training dataset.

**Disadvantages**

Random forests help for real-time prediction but that is slow in nature. It is a complex algorithm so sometimes it is hard to implement.