Google Play Apps Rating Analysis

Project focused on: Data cleaning, Exploratory Data Analysis and Machine Learning Model Building

Google Play Store Rating Predictor - Project Overview

- **Created a Machine Learning Model to estimate number of highest rating app on the Google Play Store will have.**
- **Cleaned and analyzed data provided by Kaggle.com with over 10000 apps and 13 features.**
- **Optimized Linear Regression, Gradient Boosting Regressor and Random Forest Regressor using GridsearchCV to reach the best model.**

Data Overview

Data contains 10742 rows and 13 columns: Columns (features) are:

- **App Name**
- **Category**
- **Rating**
- **Reviews**
- **Installs**
- **Size**
- **Price**
- **Type**
- **Content Rating**
- **Genres**
- **Last Updated**
- **Current Ver**
- **Android Ver**

Data Cleaning

- **I did extensive data cleaning in order to facilitate the exploratory analysis and the model building process:**
- **Fixed data scraping error and removed additional blank columns created**

- **Corrected misalignment by shifting the affected feature values to their correct column**
- **Performed data imputation based on the Category feature in order to replace null values with the closest value to the true value**
- **Dropped irrelevant features such Current Ver and Android Ve**
- **Changed target variable classes and ordered them**

Algorithms

I wanted to create a model that would make meaningful and accurate predictions for aspiring app creators to know what features are the most important when highest rating.

Split the data set into train test split of 80/20, stratified based on App.

The data would be judged based on the measure of Accuracy.

I tried 3 different models:

- **Linear Regression**
- **Random Forest Regressor**
- **Gradient Boosting Regressor**

Tools

- **Numpy and Pandas for data manipulation**
- **Scikit-learn for modeling**
- **Matplotlib and Seaborn for plotting**
- **Tableau for interactive visualizations**