

|  |
| --- |
| The purpose of this white paper is to evaluate the different types of regression models and determine which hold the lowest mean score errors. This researcher focused on several techniques that will be explained from preprocessing the data, exploratory data analysis, training/test data split, running the models, and then selecting linear and lasso for the final cross validation. The models selected for evaluation were Random Forest, Linear regression, Lasso, and Ridge regression. Ultimately, ending with a performance score for the Lasso recommended selection model.  This paper is to serve as a learning step toward creating future pipelines to evaluate all the machine learning algorithms for best model fit.  **Problem Statement:**  We live in a world where there are apps for everything. As a great idea emerges, a developer simply uploads the apps to Apple and Google stores for users to receive for free or to pay for service. Investors need predictive frameworks to predict high ratings and downloads, which can assist with determining if the app should be free, free with advertisements, or a paid service. By creating a predictive model, investors can access the mode of which the app should be uploaded to yield the strongest return on investment as the app is intended to scale.  Predicting App Ratings with regression models analysis Joseph RochelleBellevue University11/18/2020 |
|  |

**Proposal of solutions:**

By using the google play store data set, I will create three regression models: (a) un-regularized linear regression, (b) regularized linear regression, and (c) random forest regression. I will utilizes these three models in a cross-validated pipeline, by judging their performance on a data set by mean squared errors. After determining the best model, I will evaluate my final model (trained on the previous three training sets) to demonstrate an objective model that should generalize well in the future.

Using the parameters found, I will train a final model using all the data and explore the most appropriate strategy: If one of the regression models, the magnitude of the coefficients will be examined; if it is the Random Forest regression model, the feature importance will be used.

|  |  |  |
| --- | --- | --- |
| Google Play store data frame  **Pandas Profiling:**      **Data Frame:**  **The data frame that was used consisted of the google play store data set.**  **Preprocessing of the data took place which included utilizing Pandas profiling to evaluate the data frame. Missing values were replaced, values changed within the data sizing removing the “+” as numeric data, incorrect coding of row 10472 removed that did not match app naming trends, and hot encoding to convert app names to numeric values.**  **In addition to the label encoder of converting the strings, the data and time was updated to strip and match the same formats.**  **Data visualization took place to view the missing values, a heat map to examine what apps have higher predictions on average, and histograms to look at data normality.** | |  |
| **Training/Test Data Split.**  It is important to note that this researcher utilized a training and test split of the features being 'App', 'Reviews', 'Size', 'Installs', 'Type', 'Price', 'Content Rating', 'Genres', 'Last Updated', 'Current Ver' as the features the ratings as the target (e.g., X, y).  The models were evaluated using MSE to determine the errors between degrees of freedom within the bounds of unregularization and regularization. GridSearchCV was used on the selected model to determine accuracy/performance of the data.  **Recursive Feature Elimination.**  The purpose of recursive feature elimination is to examine the weights of the features. First, the estimator is trained to look at the set of features and obtain the coef\_ attributes that look at the feature importance. Then, the least important features from the current set are removed. The REFCV is a cross-validation loop of the features.  The results indicated that all features were warranted to keep within the models.  Examples of data visualization that assisted the researcher into evaluation. In addition to these graphs, the correlation matrix was reviewed to ensure no multicollinearity. This researcher utilized feature elimination within models to care for these within regularized and regularized penalties. |  | |
|  | | |
| **Model Evaluation and Selection for training.**  After cleaning the data and encoding to input into regression models, the researcher focused on evaluating each model to the other. Using consistent criteria of Mean Absolute Error, Mean Squared Error (MSE), and Root Mean Squared error, the researcher compared each score to determine which model to select. The consistent measure used was the mean squared error, which implies that the error occurs every X out of 10 observations.  MSE is an estimate construct that represents the total population within a distribution and the sum of squares by the degrees of freedom. MSE can be used to suggest that without regulation there are X errors per observation that can occur, and with regulation X errors per occur. It is with this construct that a researcher aims to make a judgement call on the selected model.  First, **Random Forest Regression model** was evaluated that utilized training data. MSE was used to compare all models as the same sampling size allowed for a review of how the errors change with penalties that are included. With Random Forrest Regression, the model seeks to predict using a forward pass at the data to learn then a backward pass, creating synapse or memory of the estimators. ***The result*** was that ***0.16 MSE*** out of 10 so a strong model, but the caution is that there could be some overfitting with the modeling. Random forest convert some categorical variables to binary and have some influence the variance that make up overfitting.  Next, **the Linear Regression model** was implemented using the training data that resulted in a **MSE of 0.25**. The linear regression modeling, by design, does not penalize for the evaluation of standard errors. This means that the model is regularized when looking at the errors and degrees of freedom. For the sake of this predictive modeling exercise, the researcher wanted to explore adding penalties to make the regression create bounds within what is considered significant.  Third, **Ridge Regression model** was performed with a **MSE of 5.59.** Some of the strongest advantages to regulation modeling is preventing overfitting which does well for ridge and lasso. Ridge specifically in this model, led to high dimensionality creating some bias as the alphas were set at p=.10. Thus, the weights of the values were pushed down and reduced overfitting but caused higher variance.  Last, the **Lasso Regression MSE of 1.97** and performed well considering the regulation of this model seeks to do feature elimination while staying within the parameter range of 0.005 to 0.3. This model also has benefits of applying as many features as needed to automatically be eliminated for which are not statistically significant within the bounds of the p=0.001 to 0.3. Moreover, the model provides the researcher with a means to test the data using GridSearchCV for the accuracy/performance of how predictable it can be.  **Conclusion**  The data set required preprocessing of the data to model it. Data was modeled using random forest, linear, ridge, and lasso regressions. These models were different in that two were unregularized that have concerns of overfitting, while the ridge and lasso were regularized adding penalties to reduce overfitting. Using sound judgement of the findings, this researcher selected the lasso model due to the feature reduction, hyperparameter controls, and scoring. To understand the accuracy/performance of the selected **Lasso model, GridSearchCV was** **utilized** with the lasso at 0.005, 0.05, 0.1, 0.2, and 0.3. The model utilized training and test data to loop through iterations to arrive at the **accuracy/performance score of 92%.** | | |