Machine Learning

CS6375.004 Final Project Report

**Google Analytics Customer Revenue Prediction**

## Team Members:

Rama Narayan Lakshmanan (rxl174430)

Shubham Raosaheb Kharde (sxk173732)

Rohit Seetepalli (rxs170631)

# 1. Introduction:

Marketing teams are challenged to make appropriate investments in promotional strategies as only small percentage of customers bring the most revenue (80%-20% rule). In this challenge, RStudio in collaboration with Google cloud and Kaggle aims to demonstrate this business impact thorough data analysis. The data that they have provided for this challenge reflects this revenue model. As part of this challenge, we are required to predict how much GStore customers will spend. GStore is an online store that sells google merchandise. The aim of this project is to train a machine learning model that best predicts revenue per customer with good metrics. We will use ensemble techniques like different types of boosting and random forests to come up with the best model. Results of the techniques with various parameters will be summarized and the best model will be presented.

# 2. Problem Definition and Algorithm:

## 2.1 Task Definition

The problem is to predict the natural log of sum of all transactions per user. It can be represented as:

The dataset that we use is from Kaggle’s Google Analytics Customer Revenue Prediction challenge [1]. There are two aspects to the challenge:

* Public leaderboard – Revenue prediction for customers during the same timeframe of 5/1/18 to 10/15/18 given in test dataset
* Private Leaderboard – Revenue prediction for customers during a future timeframe of 12/1/18 to 1/31/19.

In this project, we are solving the public leaderboard challenge.

Input to trained Model – Test.csv

Output – Prediction column with predicted log revenue

## 2.2 Algorithm Definition

The following is the algorithm that we used to get the best model. It has three main steps – preprocessing, training and testing.

* Pre-processing
  + Split train\_v2.csv into 16 different chunks so that it becomes manageable for processing
  + For each chunk,
    - Flatten all json columns – "device", "trafficSource", "geoNetwork", "totals", “customDimensions”, “hits”
    - Identify inconsistent columns between the chunks and for every column:
      * Find correlation between target and inconsistent column.
      * Drop in case of low correlation or add the column with default values.
    - Identify mixed types in features and convert them into a single data type based on the semantics of the column.
    - Resolve NaN, None and null values :
      * Replace them with default values based on their datatype
  + Concatenate all chunks into a single file.
  + Remove features that have constant values.
    - Apply filter: Number of unique values = 1
  + Parse datetime features:
    - Split datetime features that are given in POSIX format to day, month, year and hour.
    - Categorize hour into one of the following – “early morning”, “morning”, “afternoon”, “evening”, “night”, “midnight”.
  + Drop the datetime features that have been processed.
  + Parse the url features by just taking their host names and drop the processed feature.
  + Encode category data to be numeric.
  + Remove highly correlated features.
  + Scale the features using standard scaler.
  + Perform Principle Component Analysis (PCA) for further dimensionality reduction.
  + Log scale the target column.
* Training
  + Since dataset is huge, we will use the following strategy
    - Randomly sample 10 % of data
    - Fine tune the models on this data and get the top 2 best models.
    - Train and tune the top models on the whole dataset and find the best model.
  + Split the sampled dataset into training and testing in 70:30 ratio.
  + Random Forest Model
    - Build a PARAM grid to tune the model with following parameters in a 5 fold cross validation.
      * No of Trees.
      * Max depth
    - Use cross validator to find the best model with the algorithm’s default scoring metric.
  + LGBM:
    - Build a PARAM grid to tune the model with following parameters in a 5 fold cross validation.
      * Regularization Parameter
      * Elastic Net Parameter
      * Max Iterations
    - Use cross validator to find the best model with the algorithm’s default scoring metric.
  + Gradient boosting:
    - Build a PARAM grid to tune the model with following parameters in a 5 fold cross validation.
      * Regularization Parameter
      * Elastic Net Parameter
      * Max Iterations
    - Use cross validator to find the best model with the algorithm’s default scoring metric.
  + Adaboost:
    - Build a PARAM grid to tune the model with following parameters in a 5 fold cross validation.
      * Regularization Parameter
      * Elastic Net Parameter
      * Max Iterations
    - Use cross validator to find the best model with the algorithm’s default scoring metric.
  + XGBoost:
    - Build a PARAM grid to tune the model with following parameters in a 5 fold cross validation.
      * Regularization Parameter
      * Elastic Net Parameter
      * Max Iterations
    - Use cross validator to find the best model with the algorithm’s default scoring metric.
  + Find top 2 best models.
* Testing
  + Regression Metrics:
    - RMSE – Root Mean Square error
    - R-Squared to determine the goodness of the fit
    - MAE – Mean Absolute Error
    - Explained Variance
    - MSE –Mean Square Error
  + Use the top models on the testing data.
  + We have predicted the log revenue. To get the competition metric do the following:
    - Inverse log the prediction column
    - Group the prediction by “fullVisitorId”
    - Apply log again
  + Get the regression metrics on this data.
  + Based on the metrics obtained, decide on the best model. Our goal is to choose a model with highest R-Squared and least RMSE value.

As this an ongoing competition, we list out the steps on how to prepare and upload the results to Kaggle.

* Run the preprocessing on test.csv
* Run the chosen best model on the preprocessed test data.
* Select only “fullVisitorId” column and “prediction” column.
* Export it as csv and upload it to Kaggle website.

# 3. Technology Used

* Python
* Scikit-learn library
* MS Excel for data exploration

# 4. Experimental Evaluation

## 4.1 Methodology

### 4.1.1 Dataset Description

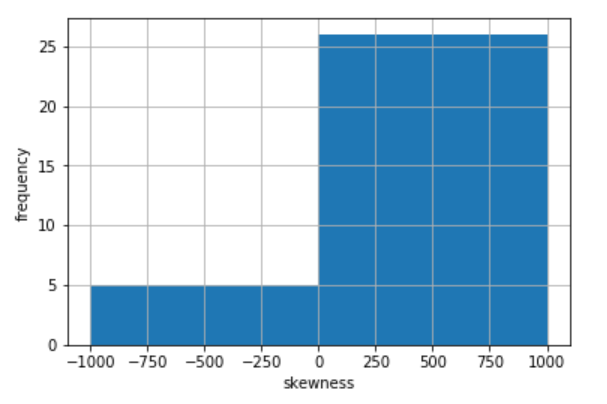
There are two datasets provided:

* train\_v2.csv
  + No of records – 1,708,344
  + No of features – 14, 6 of which are deep nested json blobs which we will expand during preprocessing.
  + Label column – transactionRevenue nested within “totals” feature.
* test\_v2.csv
  + No of records – 401,589
  + No of features – 14, 6 of which are deep nested json blobs which we will expand during preprocessing.

We have two major challenges that makes this dataset interesting. The first challenge is to flatten the deep nested json blob features. The second challenge is the humongous training data and right skewed target data. We have 1,708,344 records of which 1,689,830 records have transaction revenue of 0 which is 98.91 % of data. This is in accordance with the theme of this challenge that smaller number of people bring more revenue.

#### Data Distribution:

There are many non-numeric columns. Before pre-processing, we just had a look at the skewness of numeric columns. Some of them are left skewed and some are right skewed as observed in the below histogram.



### 4.1.2 Data Preprocessing

#### Data Flattening:

There are six json blobs to flatten:

* customDimension
* device
* geoNetwork
* hits
* totals
* trafficSource

##### customDimesion

Has index and value. Flatten it to customDimensions\_index and customDimensions\_value. We are not interested in index. So drop it.

##### Device

Flatten it to device\_{attributes}

##### geoNetwork

Flatten it to geoNetwork\_{attributes}

##### hits

Flatten it to hits\_{attributes}. Drop the json blobs that remain after flattening.

##### totals

Flatten it to totals\_{attributes}. “totals\_transactionRevenue” is the target column

##### trafficSource

Flatten it to trafficSource\_{attributes}. There is one more blob “trafficSource\_adwordsClickInfo”. Flatten it again to trafficSource\_adwordsClickInfo\_{attributes}

#### Feature Engineering

##### Inconsistent features

Since we are processing in chunks, not all chunks have all the features. Identify the inconsistent features and apply the following steps for each inconsistent column.

* Find correlation between inconsistent column and target
* Drop if the correlation is very less or add the column with default value in the chunks that doesn’t have them.

At the end of this process, total features – 120 (excluding target column)

##### Constant Features

Drop all the features which have a constant value for all records. The features with constant values are:

*['device\_browserVersion', 'device\_browserSize', 'device\_operatingSystemVersion', 'device\_mobileDeviceBranding', 'device\_mobileDeviceModel', 'device\_mobileInputSelector', 'device\_mobileDeviceInfo', 'device\_mobileDeviceMarketingName', 'device\_flashVersion', 'device\_language', 'device\_screenColors', 'device\_screenResolution', 'geoNetwork\_cityId', 'geoNetwork\_latitude', 'geoNetwork\_longitude', 'geoNetwork\_networkLocation', 'trafficSource\_adwordsClickInfo\_criteriaParameters', 'hits\_appInfo.screenDepth', 'hits\_contentGroup.contentGroup4', 'hits\_contentGroup.contentGroup5', 'hits\_page.searchCategory', 'hits\_time']*

Total Features Dropped: 22

Remaining features: 98

##### Incorrect Features

Drop the following features

*'hits\_social.socialInteractionNetworkAction'* - has only “:” and “unknown” as values

*trafficSource\_adwordsClickInfo\_targetingCriteria* - has only “{}” and “unknown” as values

Total Features Dropped: 2

Remaining features: 96

##### Redundant Features

Drop the following redundant features:

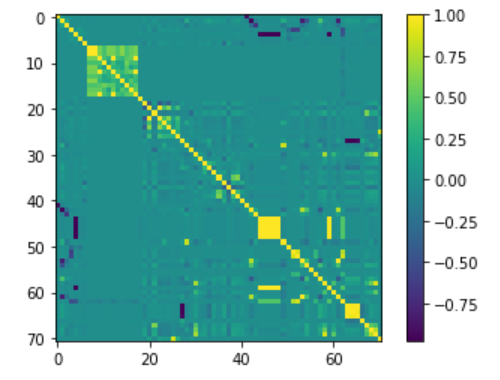
* *"customDimensions\_index"* – It’s key for another column *"customDimensions\_value"*
* *"hits\_page.pagePath" , "hits\_page.pagePathLevel1" , "hits\_page.pagePathLevel2" ,"hits\_page.pagePathLevel3","hits\_page.pagePathLevel4"* – All this information about the page is aggregated in *“hits\_page.pageTitle”*
* *“trafficSource\_referralPath”* - information is covered in *“trafficSource\_source”*
* *“hits\_appInfo.screenName*” - same as *“hits\_appInfo.landingScreenName”*
* “*trafficSource\_adwordsClickInfo\_gclId*” - not required as its unique id for all clicks.

Total Features Dropped: 9

Remaining features: 87

##### Correlated Features

We found pearson correlation between the features and dropped features that have a correlation of more than 0.8 for positive correlation and less than -0.8 for negative correlation.



Darker shade of yellow indicates high positive correlation and darker shade of blue indicates high negative correlation.

Total Features Dropped: 17

Remaining features: 70

##### Feature addition

###### Date Features

Split the date features as month, day and year. Drop the date feature once you split

Date feature – “date”

New features added – 3

Features dropped – 1

Total features - 72

###### DateTime Features

Split the datetime features as month, day, year and hour in UTC. Categorize hour as

* 12am – 3.59am – Midnight
* 4am - 6.59am – EarlyMorning
* 7am – 11.59am – Morning
* 12pm – 3.59pm – Afternoon
* 4pm – 6.59pm – evening
* 7pm – 11.59pm – night

Date feature – “visitStartTime”

New features added – 4

Features dropped – 1

Total features – 75

###### Additional Features

We added one additional feature.

*“isEntryExitSame”* – Boolean to determine if the landing screen the same as the exit screen of the user’s visit

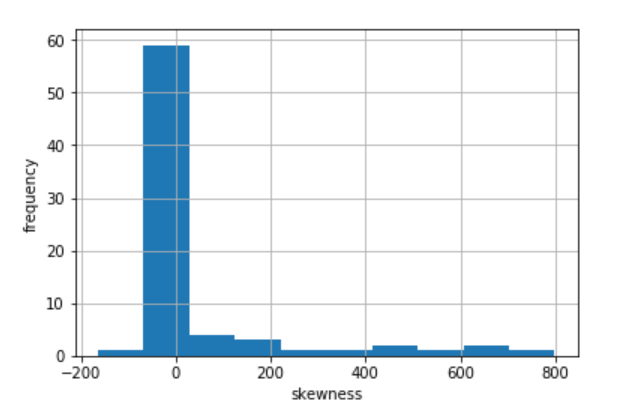
Total Features after feature engineering – 76

###### Final Feature Description

|  |  |
| --- | --- |
| Features | Description |
| channelGrouping | The Default Channel Group associated with an end user's session for this View. |
| customDimensions\_value | The value of the custom dimension. |
| date\_day | date in day |
| date\_month | date in month |
| date\_year | date in year |
| device\_browser | The browser used (e.g., "Chrome" or "Firefox"). |
| device\_deviceCategory | The type of device (Mobile, Tablet, Desktop). |
| device\_isMobile | If the user is on a mobile device, this value is true, otherwise false. |
| device\_operatingSystem | The operating system of the device (e.g., "Macintosh" or "Windows"). |
| fullVisitorId | The unique visitor ID (also known as client ID). |
| geoNetwork\_city | Users' city, derived from their IP addresses or Geographical IDs. |
| geoNetwork\_continent | The continent from which sessions originated, based on IP address. |
| geoNetwork\_country | The country from which sessions originated, based on IP address. |
| geoNetwork\_metro | The Designated Market Area (DMA) from which sessions originate. |
| geoNetwork\_networkDomain | The domain name of user's ISP, derived from the domain name registered to the ISP's IP address. |
| geoNetwork\_region | The region from which sessions originate, derived from IP addresses. In the U.S., a region is a state, such as New York. |
| geoNetwork\_subContinent | The sub-continent from which sessions originated, based on IP address of the visitor. |
| hits\_appInfo.exitScreenName | The exit screen of the session. |
| hits\_appInfo.landingScreenName | The landing screen of the session. |
| hits\_contentGroup.contentGroup1 | The content group on a property. A content group is a collection of content that provides a logical structure that can be determined by tracking-code or page-title/URL regex match, or predefined rules. |
| hits\_contentGroup.contentGroup3 | The content group on a property. A content group is a collection of content that provides a logical structure that can be determined by tracking-code or page-title/URL regex match, or predefined rules. |
| hits\_contentGroup.contentGroupUniqueViews3 | number of unique content group views |
| hits\_dataSource | The data source of a hit. By default, hits sent from analytics.js are reported as "web" and hits sent from the mobile SDKs are reported as "app". |
| hits\_eCommerceAction.action\_type | action type |
| hits\_eCommerceAction.option | This field is populated when a checkout option is specified. For example, a shipping option such as option = 'Fedex'. |
| hits\_eCommerceAction.step | This field is populated when a checkout step is specified with the hit. |
| hits\_eventInfo.eventCategory | The event category. |
| hits\_eventInfo.eventLabel | The event label. |
| hits\_exceptionInfo.isFatal | If the exception was fatal, this is set to true. |
| hits\_hitNumber | The sequenced hit number. For the first hit of each session, this is set to 1. |
| hits\_hour | The hour in which the hit occurred (0 to 23). |
| hits\_isEntrance | If this hit was the first pageview or screenview hit of a session, this is set to true. |
| hits\_isExit | If this hit was the last pageview or screenview hit of a session, this is set to true. |
| hits\_isInteraction | If this hit was an interaction, this is set to true. If this was a non-interaction hit (i.e., an event with interaction set to false), this is false. |
| hits\_item.currencyCode | The local currency code for the transaction. |
| hits\_latencyTracking.domainLookupTime | The total time (in milliseconds) all samples spent in DNS lookup for this page. |
| hits\_latencyTracking.domInteractiveTime | The time (in milliseconds), including the network time from users' locations to the site's server, the browser takes to parse the document (DOMInteractive). |
| hits\_latencyTracking.domLatencyMetricsSample | Sample set (or count) of pageviews used to calculate the averages for site speed DOM metrics. |
| hits\_latencyTracking.pageDownloadTime | The total time (in milliseconds) to download this page among all samples. |
| hits\_latencyTracking.pageLoadSample | The sample set (or count) of pageviews used to calculate the average page load time. |
| hits\_latencyTracking.pageLoadTime | Total time (in milliseconds), from pageview initiation (e.g., a click on a page link) to page load completion in the browser, the pages in the sample set take to load. |
| hits\_latencyTracking.redirectionTime | The total time (in milliseconds) all samples spent in redirects before fetching this page. If there are no redirects, this is 0. |
| hits\_latencyTracking.serverConnectionTime | Total time (in milliseconds) all samples spent in establishing a TCP connection to this page |
| hits\_latencyTracking.serverResponseTime | The total time (in milliseconds) the site's server takes to respond to users' requests among all samples; this includes the network time from users' locations to the server. |
| hits\_latencyTracking.speedMetricsSample | The sample set (or count) of pageviews used to calculate the averages of site speed metrics. |
| hits\_minute | The minute in which the hit occurred (0 to 59). |
| hits\_page.hostname | The hostname of the URL. |
| hits\_page.pageTitle | The page title. |
| hits\_page.searchKeyword | If this was a search results page, this is the keyword entered. |
| hits\_promotionActionInfo.promoIsClick | True if the Enhanced Ecommerce action is a promo click. |
| hits\_promotionActionInfo.promoIsView | True if the Enhanced Ecommerce action is a promo view. |
| hits\_referer | The referring page, if the session has a goal completion or transaction. If this page is from the same domain, this is blank. |
| hits\_social.hasSocialSourceReferral | A string, either Yes or No, that indicates whether sessions to the property are from a social source. |
| hits\_social.socialNetwork | The social network name. This is related to the referring social network for traffic sources; e.g., Google+, Blogger. |
| isEntryExitSame | Is landing screen same as exit screen |
| totals\_bounces | Total bounces (for convenience). For a bounced session, the value is 1, otherwise it is null. |
| totals\_hits | Total number of hits within the session. |
| totals\_newVisits | Total number of new users in session (for convenience). If this is the first visit, this value is 1, otherwise it is null. |
| totals\_sessionQualityDim | An estimate of how close a particular session was to transacting, ranging from 1 to 100, calculated for each session |
| totals\_timeOnSite | Total time of the session expressed in seconds. |
| totals\_totalTransactionRevenue | Total transaction revenue, expressed as the value passed to Analytics multiplied by 10^6 |
| totals\_transactions | Total number of ecommerce transactions within the session. |
| trafficSource\_adContent | The ad content of the traffic source. Can be set by the utm\_content URL parameter. |
| trafficSource\_adwordsClickInfo\_isVideoAd | True if it is a Trueview video ad. |
| trafficSource\_adwordsClickInfo\_targetingCriteria | Google Ads targeting criteria for a click. There are multiple types of targeting criteria, but should have only one value for each criterion. |
| trafficSource\_campaign | The campaign value. Usually set by the utm\_campaign URL parameter. |
| trafficSource\_isTrueDirect | True if the source of the session was Direct |
| trafficSource\_keyword | The keyword of the traffic source, usually set when the trafficSource.medium is "organic" or "cpc". Can be set by the utm\_term URL parameter. |
| trafficSource\_medium | The medium of the traffic source. Could be "organic", "cpc", "referral", or the value of the utm\_medium URL parameter. |
| trafficSource\_source | The source of the traffic source. Could be the name of the search engine, the referring hostname, or a value of the utm\_source URL parameter. |
| visitId | An identifier for this session |
| visitNumber | The session number for this user. If this is the first session, then this is set to 1. |
| visitStartTime\_day | Visit start time in day |
| visitStartTime\_hour | Visit start time in hour |
| visitStartTime\_month | Visit start time in month |
| visitStartTime\_year | Visit start time in year |

Table content reference - <https://support.google.com/analytics/answer/3437719?hl=en>

#### Scaling:

We encoded all category data except “fullVisitorId” (Id column) to numeric and calculated skewness. Some features are right skewed and some are left skewed as observed in the following histogram.

We performed standard scaling on the data.

## 4.1.3 Algorithm Selection and Model Evaluation

### Regression Techniques used

We tried the following ensemble techniques

* Random Forests
* Gradient Boosting
* Adaboost
* LGBM
* XGBoost

### Model evaluation metrics

*RMSE =*

*MAE =*

*R-Squared =*

*Where SSE =*

*SSTO =*

*Explained Variance =*

*MSE =*

## 4.2 Experimental Results and Analysis

### 4.2.1 Linear Regression

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Linear Regression Hyper-Parameter Tuning | | | Parameter Grid | Best Model Value | |
| Run 1 | | maxIter | 100, 400, 800 | 100 | |
| regParam | 0.001, 0.01, 0.1 | 0.1 | |
| elasticNetParam | 0.0,0.5,1.0 | 1.0 | |
|  | | | | | |
| Run 2 | | maxIter | 50,75,100 | 50 | |
| regParam | 0.1,0.3,0.5,1.0 | 1.0 | |
| elasticNetParam | 0.0,0.5,1.0 | 1.0 | |
| **The above results suggest that lower iterations, greater regularization parameter and 1.0 elastic net parameter (Lasso Model) yield better results** | | | | | |
| LR Parameters | MSLE | RMSLE | MALE | EX Log Var | R-Squared |
| PCA-20  Iter-40  Reg-0.5  Elatstic-1  spear-0.08  Folds-3 | 3.8457845048126393 | 1.961067185185821 | 1.575106221069996 | 0.7737998582637693 | -0.08060837605711568 |
| PCA-20  Iter-40  Reg-0.5  Elatstic-1  spear-0.07  Folds-3 | 3.4778001464178514 | 1.8648860947569563 | 1.5067391146465643 | 0.77366046547347 | -0.010420821804575597 |
| No PCA  Iter-40  Reg-0.5  Elatstic-1  spear-0.07  Folds-3 | 7.58685512696336 | 2.7544246453594186 | 1.7372273395611544 | 6.474253222869448 | 0.0330832760423867 |
| **The above readings suggest that PCA generates negative R-Square. Running same parameters without PCA give better results.** | | | | | |
| No PCA  Iter-40  Reg-0.01  Elatstic-1  spear-0.07  Folds-3 | 7.586853116533982 | 2.754424280413964 | 1.737221820849303 | 6.474291843821258 | 0.03308314062356532 |
| No PCA  Iter-40  Reg-0.5  Elatstic-1  spear-0.07  Folds-3 | 7.58685512696336 | 2.7544246453594186 | 1.7372273395611544 | 6.474253222869448 | 0.0330832760423867 |
| No PCA  Iter-40  Reg-5  Elatstic-1  spear-0.07  Folds-3 | 7.674130980805224 | 2.770222189790058 | 1.7677356143350529 | 6.32974912544915 | 0.028417428462864436 |
| No PCA  Iter-40  Reg-10  Elatstic-1  spear-0.07  Folds-3 | 7.715938053129935 | 2.7777577383799934 | 1.7562029199165714 | 6.527694885705374 | 0.015719871987001754 |
| No PCA  Iter-40  Reg-20  Elatstic-1  spear-0.07  Folds-3 | 7.7159116810992385 | 2.777752991376166 | 1.7565036548866935 | 6.525464712843575 | 0.015723057719717826 |
| **The above readings suggest that the regularization parameter yield better results for 0.5 value.** | | | | | |
| No PCA  Iter-40  Reg-0.5  Elatstic-1  spear-0.06  Folds-3 | 14.819393284379434 | 3.8495965092954134 | 2.0726358538379706 | 11.754807932377721 | 0.046334213327464346 |
| No PCA  Iter-40  Reg-0.5  Elatstic-1  spear-0.07  Folds-3 | 7.58685512696336 | 2.7544246453594186 | 1.7372273395611544 | 6.474253222869448 | 0.0330832760423867 |
| No PCA  Iter-40  Reg-0.5  Elatstic-1  spear-0.08  Folds-3 | 7.74980944329994 | 2.78384795621096 | 1.806135992171305 | 6.261368046324617 | -0.062453519494334975 |
| No PCA  Iter-40  Reg-0.5  Elatstic-1  spear-0.09  Folds-3 | 3.948501999201379 | 1.9870837926975748 | 1.5899320354575388 | 0.8201328202415203 | -0.10330326403998247 |
| **The above results suggest that the spearman coefficient threshold of 0.007 yields better results. Even though R-square value for 0.006 is better, its RMSLE is worst.** | | | | | |
| No PCA  Iter-20  Reg-0.5  Elatstic-1  spear-0.07  Folds-3 | 7.852906311735545 | 2.802303750797823 | 1.8092824124928548 | 6.2317747557536265 | 0.03518688927702818 |
| No PCA  Iter-40  Reg-0.5  Elatstic-1  spear-0.07  Folds-3 | 7.58685512696336 | 2.7544246453594186 | 1.7372273395611544 | 6.474253222869448 | 0.0330832760423867 |
| No PCA  Iter-80  Reg-0.5  Elatstic-1  spear-0.07  Folds-3 | 14.4617112523787 | 3.802855670726763 | 2.1028257237094525 | 11.739840422160565 | 0.03117928633305611 |
| **The above results suggest that lower the iterations, better the results** | | | | | |
| No PCA  Iter-10000  Reg-0.5  Elatstic-1  spear-0.07  Folds-3 | 7.742136153255876 | 2.7824694343794465 | 1.7809443552771569 | 6.379484385791095 | 0.03830516991845867 |
| No PCA  Iter-10000  Reg-0.001  Elatstic-1  spear-0.07  Folds-3 | 7.729283982885915 | 2.7801589851815875 | 1.7780365125199389 | 6.384456613267146 | 0.038685931757443615 |
| **Increasing the iterations or decreasing the regularization parameter have negligible effect on result.** | | | | | |
| No PCA  Iter-10  Reg-1  Elatstic-1  pear-0.15  Folds-10 | 4.212309897540613 | 2.0523912632684373 | 1.589843355268758 | 1.340677201588615 | 0.06681874923575615 |
| **The above result is the best that we achieved with Linear Regression.**  **R-Square is very low and hence we conclude that Linear Regression is not the best choice of model for this dataset.** | | | | | |

### 4.2.2 Random Forest Regression

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Random Forest Hyper-Parameter Tuning | | | Parameter Grid | Best Model Value | |
| Parameters | | numTrees | 5, 10, 20 | 20 | |
| maxDepth | 2, 4, 6 | 6 | |
| **The above results suggest that more the number of trees and depth, better will be the results.** | | | | | |
| RF Parameters | Kaggle RMSLE | Competition Score Screenshot | | | |
| PCA-20  trees-100  depth-10  spear-0.07  Folds-3 | 2.02 |  | | | |
| No PCA  trees-100  depth-10  spear-0.07  Folds-3 | 1.93 |  | | | |
| **The above results suggest that better results are produced without using PCA dimensionality reduction** | | | | | |
| RF Parameters | Kaggle RMSLE | Competition Score Screenshot | | | |
| No PCA  trees-100  depth-10  spear-0.07  Folds-3 | 1.93 |  | | | |
| PCA-15  trees-100  depth-10  pear-0.15  Folds-3 | 1.91 |  | | | |
| **The above results suggest that anyone of the Spearman (0.007 threshold - 277 features) and Pearson correlation (0.15 threshold - 268 features) techniques can be chosen for the feature reduction.** | | | | | |
| RF Parameters | MSLE | RMSLE | MALE | EX Log Var | R-Squared |
| No PCA  trees-50  depth-5  spear-0.07  Folds-3 | 3.6362268280233203 | 1.90688930670433 | 1.4976251046584386 | 0.6030258293651772 | -0.052877585173111585 |
| No PCA  trees-50  depth-10  spear-0.07  Folds-3 | 3.3422710445714814 | 1.8281879128173562 | 1.4456274200770036 | 0.5766703225968669 | -0.030397629643458712 |
| No PCA  trees-50  depth-15  spear-0.07  Folds-3 | 3.1203910966525275 | 1.7664628772358981 | 1.413792684487838 | 0.569728312570852 | -0.029602069529155006 |
| No PCA  trees-50  depth-20  spear-0.07  Folds-3 | 2.967994131592468 | 1.7227867342165333 | 1.377592052471503 | 0.5805144306711325 | 0.014041499762836751 |
| **The above results suggest that increasing the depth improves the metrics.** | | | | | |
| No PCA  trees-50  depth-30  spear-0.07  Folds-3 | 2.884523021700081 | 1.6983883600932035 | 1.3657913060295788 | 0.6533175290047863 | 0.0024321647111059663 |
| No PCA  trees-250  depth-30  pear-0.2  Folds-5 | 2.5134432514740834 | 1.5853842598796304 | 1.2291520761445962 | 0.9786449218927207 | 0.3015627631242761 |
| **The above results suggest that increasing the number of trees improves the metrics.** | | | | | |
| No PCA  trees-250  depth-30  pear-0.2  Folds-5 | MSLE  2.5134432514740834 | RMSLE  1.5853842598796304 | MALE  1.2291520761445962 | Ex Log Variance  0.9786449218927207 | R-Squared =  0.3015627631242761 |
| **The above result is the best for Random Forest regressor. It also the best among both models.**  **Following is the highest score screenshot that we achieved in the competition using this model.** | | | | | |
|  | | | | | |

# 5. Future Work

In this project, we restricted our techniques to the algorithms that are available in Spark MLlib. We recommend trying other ML techniques to increase the R-square. In addition, we recommend trying SVD for dimensionality reduction.

# 6. Conclusion

Feature engineering was important as we had more features than records. Getting rid of the features having zero values, performing spearman correlation to select the important features and scaling the features were part of the feature engineering we did. We tried PCA for dimensionality reduction but had to disregard it as it didn’t give better results. Of all the models we ran, random forest model gave better results in terms of RMSLE. We scored 1.81 RSMLE in the Kaggle competition public leaderboard using Random Forest Model.

# 7. Contribution of team members

|  |  |
| --- | --- |
| **RAMA NARAYAN LAKSHMANAN** | **SHUBHAM RAOSAHEB KHARDE** |
| Initial Data Analysis | Correlation |
| Scaling | Random Forest Technique |
| Linear Regression Technique | Result aggregation |
| Project Report | Project Report |
| Source Code Preparation | Kaggle output preparation |

# 8. References

[1] <https://www.kaggle.com/c/santander-value-prediction-challenge>

<https://spark.apache.org/docs/2.3.0/mllib-evaluation-metrics.html>

<https://spark.apache.org/docs/latest/mllib-guide.html>

Stack overflow for fixing errors