Machine Learning Project

SJSU SAN JOSÉ STATE UNIVERSITY

Team 3:

Aditi Khurd, Ashna Gupta

Koushik Kumar Kamala, Madhusudhan Shagam

(KAGGLE COMPETITION) PROBLEM: Google Analytics Customer Revenue Prediction

- To analyze a Google Merchandise Store(also known as GStore where Google swag is sold) customer dataset to predict revenue per customer

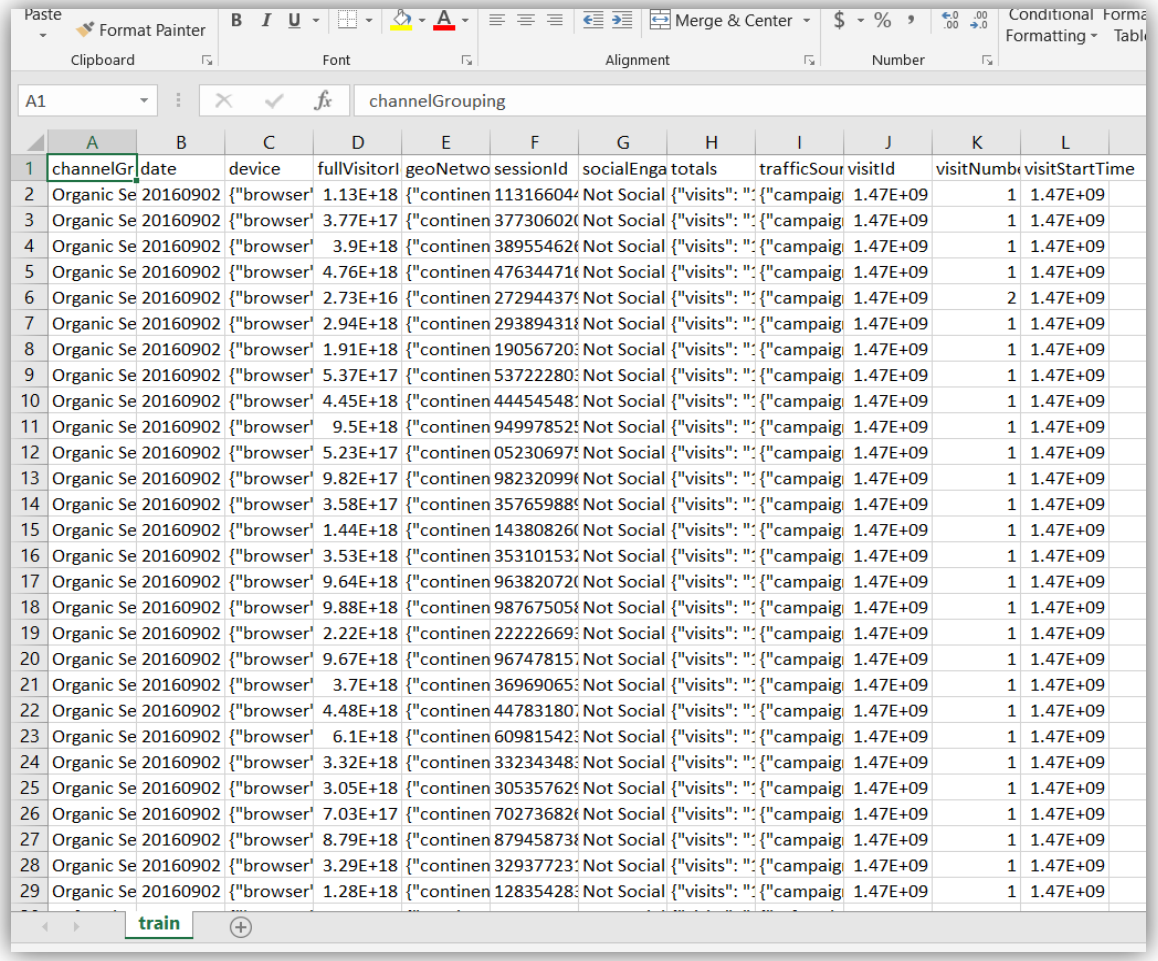
MOTIVATION

- Kaggle competition provided us an opportunity to explore full scale machine learning problems.
- The predicted revenue per customer can be used to improve the marketing budgets of companies relying on this data.

DataSet

- Source of Dataset: Kaggle.
- Quantitatively it is 9 lakhs with 13 columns.
- It is a .csv file included with data of JSON format.
- Extracted 55 features after decoding JSON data.

Data Shared: Training & Test



Paste Format Painter | Clipboard | Font | Alignment | Merge & Center | Number | Conditional Formatting

A1 channelGrouping

	A	B	C	D	E	F	G	H	I	J	K	L
1	channelGr	date	device	fullVisitorI	geoNetwo	sessionId	socialEnga	totals	trafficSour	visitId	visitNumb	visitStartTime
2	Organic Se	20160902	{"browser"	1.13E+18	{"continen	11316604	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
3	Organic Se	20160902	{"browser"	3.77E+17	{"continen	37730602	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
4	Organic Se	20160902	{"browser"	3.9E+18	{"continen	38955462	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
5	Organic Se	20160902	{"browser"	4.76E+18	{"continen	47634471	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
6	Organic Se	20160902	{"browser"	2.73E+16	{"continen	27294437	Not Social	{"visits": "	{"campaig	1.47E+09	2	1.47E+09
7	Organic Se	20160902	{"browser"	2.94E+18	{"continen	29389431	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
8	Organic Se	20160902	{"browser"	1.91E+18	{"continen	19056720	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
9	Organic Se	20160902	{"browser"	5.37E+17	{"continen	53722280	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
10	Organic Se	20160902	{"browser"	4.45E+18	{"continen	44454548	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
11	Organic Se	20160902	{"browser"	9.5E+18	{"continen	94997852	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
12	Organic Se	20160902	{"browser"	5.23E+17	{"continen	05230697	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
13	Organic Se	20160902	{"browser"	9.82E+17	{"continen	98232099	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
14	Organic Se	20160902	{"browser"	3.58E+17	{"continen	35765988	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
15	Organic Se	20160902	{"browser"	1.44E+18	{"continen	14380826	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
16	Organic Se	20160902	{"browser"	3.53E+18	{"continen	35310153	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
17	Organic Se	20160902	{"browser"	9.64E+18	{"continen	96382072	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
18	Organic Se	20160902	{"browser"	9.88E+18	{"continen	98767505	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
19	Organic Se	20160902	{"browser"	2.22E+18	{"continen	22222669	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
20	Organic Se	20160902	{"browser"	9.67E+18	{"continen	96747815	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
21	Organic Se	20160902	{"browser"	3.7E+18	{"continen	36969065	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
22	Organic Se	20160902	{"browser"	4.48E+18	{"continen	44783180	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
23	Organic Se	20160902	{"browser"	6.1E+18	{"continen	60981542	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
24	Organic Se	20160902	{"browser"	3.32E+18	{"continen	33234348	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
25	Organic Se	20160902	{"browser"	3.05E+18	{"continen	30535762	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
26	Organic Se	20160902	{"browser"	7.03E+17	{"continen	70273682	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
27	Organic Se	20160902	{"browser"	8.79E+18	{"continen	87945873	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
28	Organic Se	20160902	{"browser"	3.29E+18	{"continen	32937723	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09
29	Organic Se	20160902	{"browser"	1.28E+18	{"continen	12835428	Not Social	{"visits": "	{"campaig	1.47E+09	1	1.47E+09

train

Task:

Regression Model

Training Set:

9lakh rows, 13 columns

Test Set:

7lakh rows, 9 columns

Predictors:

Boosting Algorithm and RF

What's good about Gstore data:

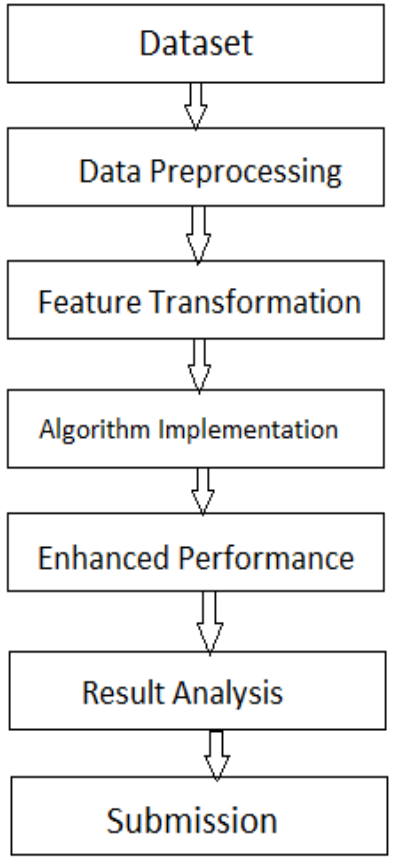
-Daily update of data based on their Merchandise.

-Raw data consists of missing values.

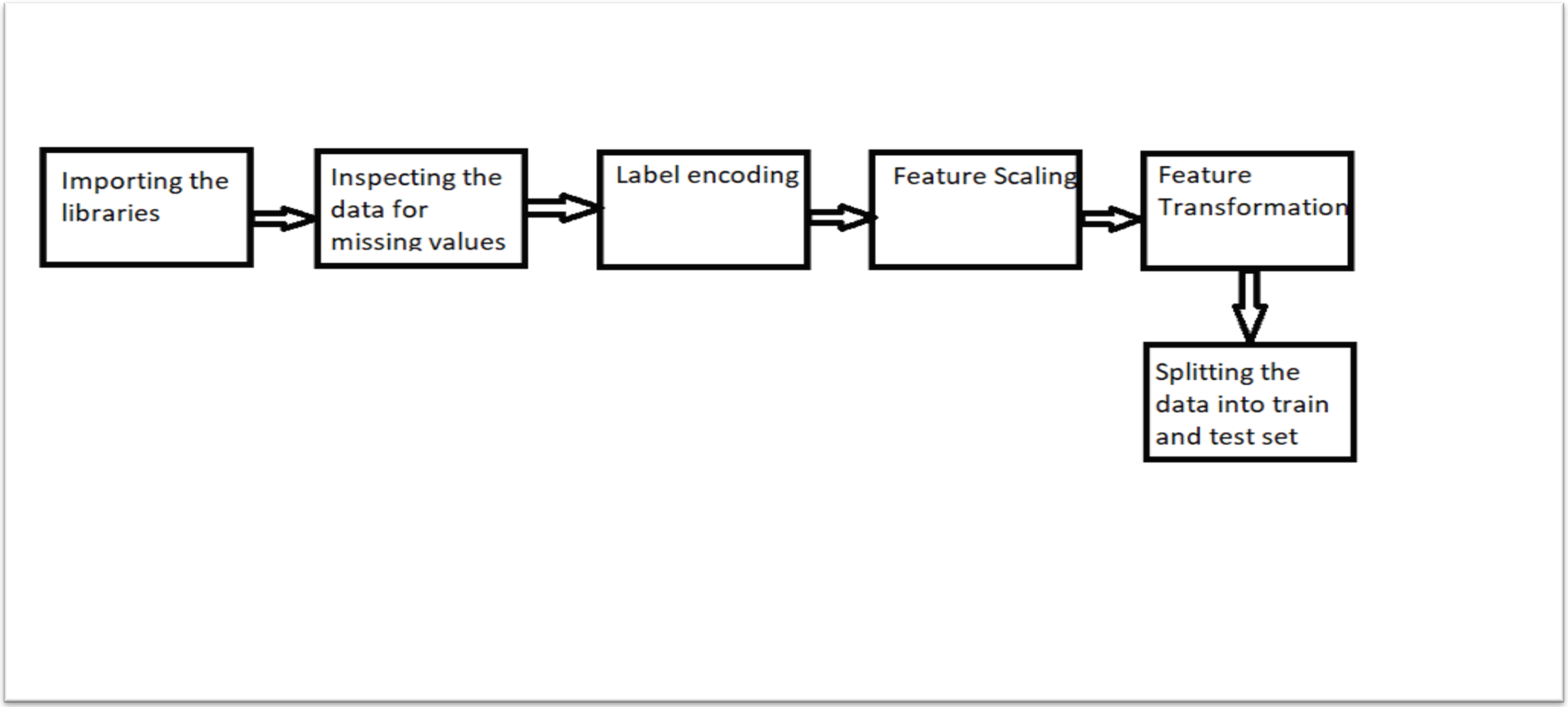
-Large Dataset with multiple features

WORKFLOW

- Took dataset from Kaggle.
- Conversion of data from JSON to CSV.
- Clean the missing data and conversion of date datatype
- Feature scaling.
- Label encoding.
- Removal of features not included in test dataset.
- Cross-fold Validation.
- Implementation of various algorithms.
- Metrics
- Result analysis and graphs.



DATA PREPROCESSING



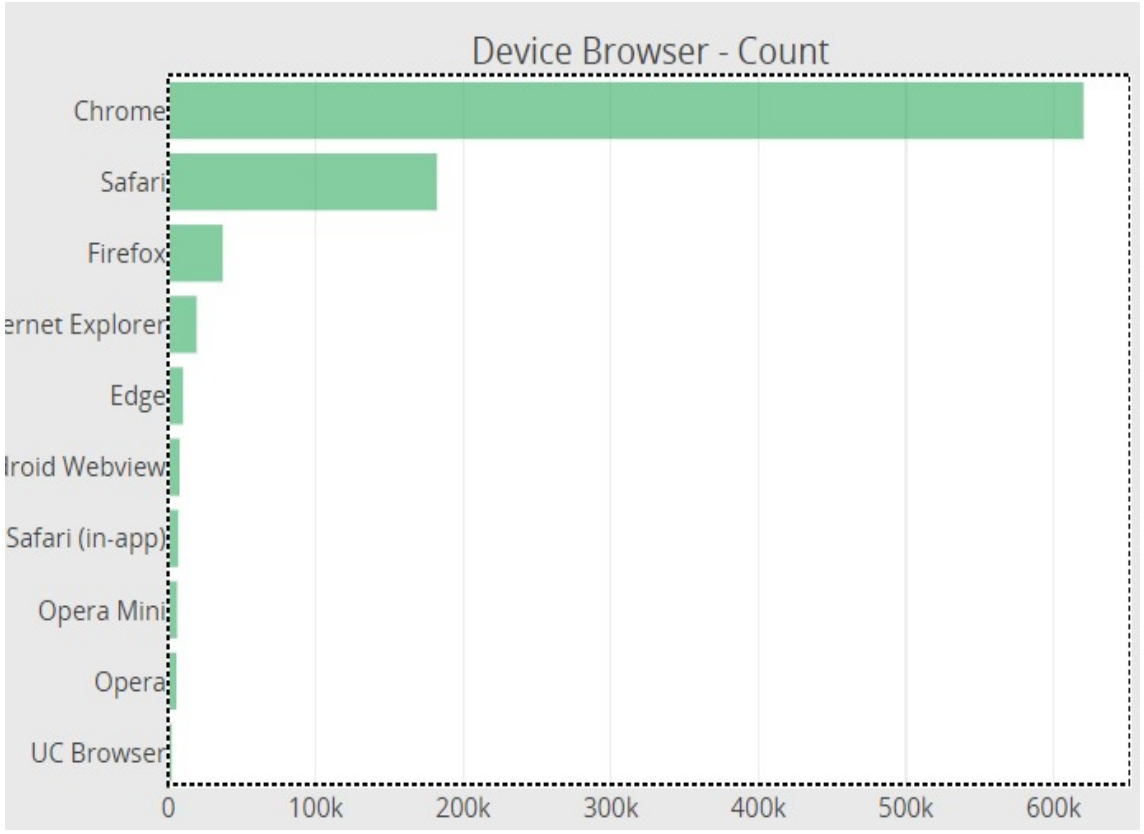
Preprocessing

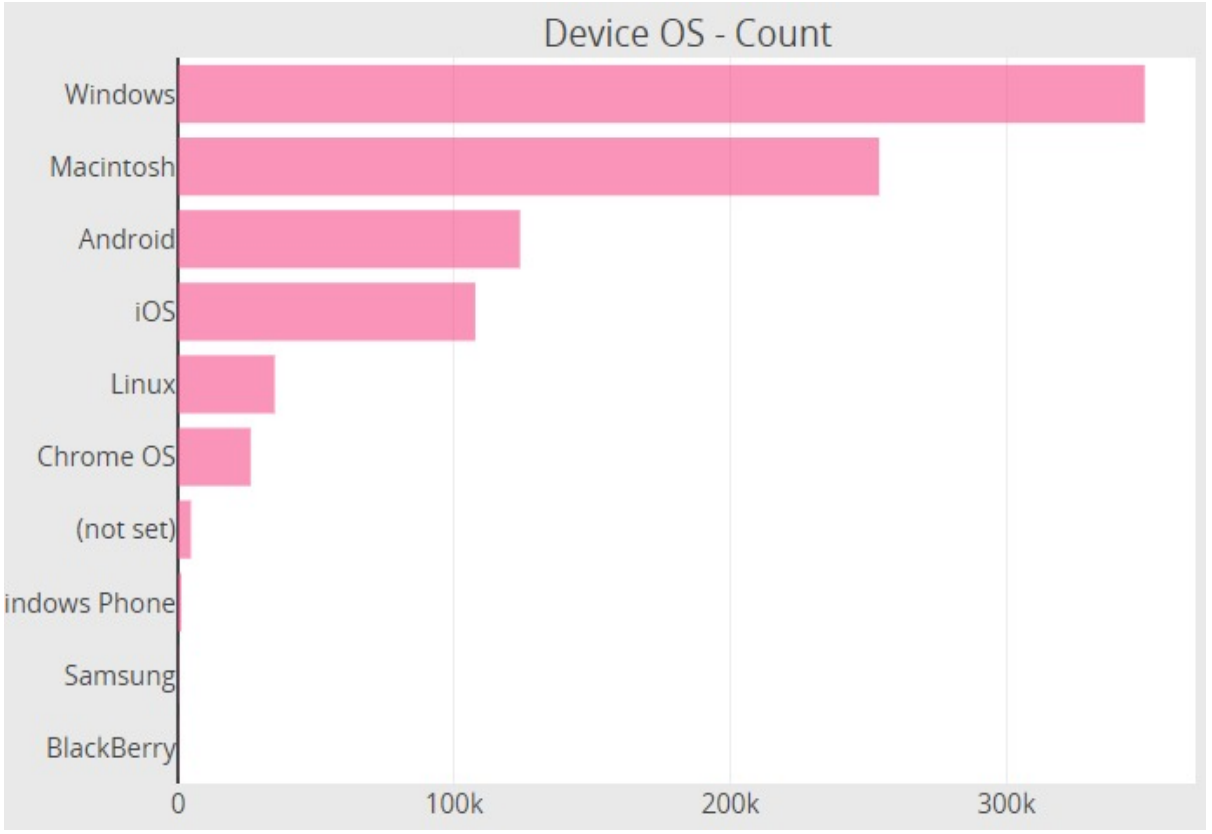
- Importing the libraries: NumPy, Pandas, SK Learn, Seaborn, Matplotlib
- Inspect the dataset and operate on missing values using SK learn .
- Perform Label Encoding and convert numerical data to floating point values.
- Split the data into trainset and test set based on the dates.
- Feature Scaling.
- Removal of features which are not included in the test set.

Insights from preprocessing

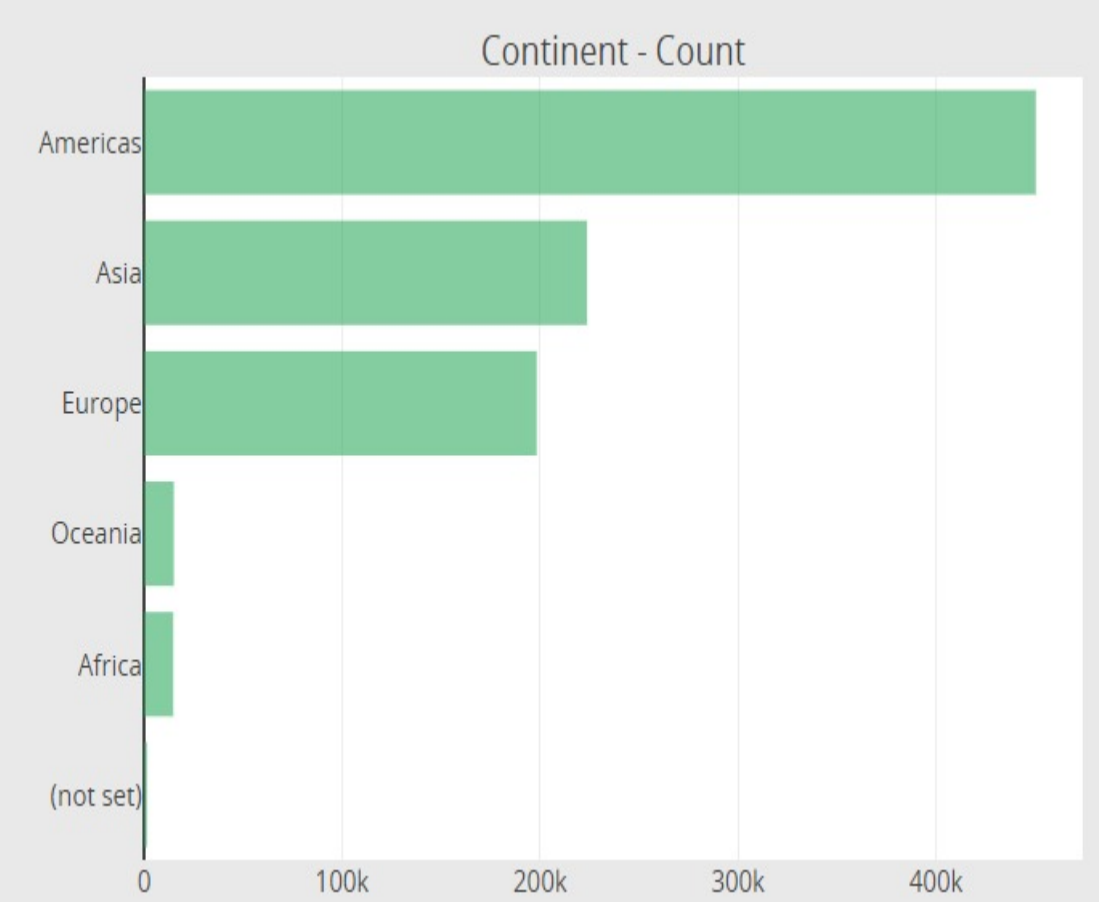
- Non-zero revenue was observed for only 1.3% users.
- Most frequently used browser for generating the revenue is observed to be chrome.
- Desktop is the most popular device used by gstore customers.
- Most used operating system is Windows OS.
- We could conclude that America is the most popular wherein Northern America is the most revenue generated subcontinent.
- Revenue is also generated from unregistered users.

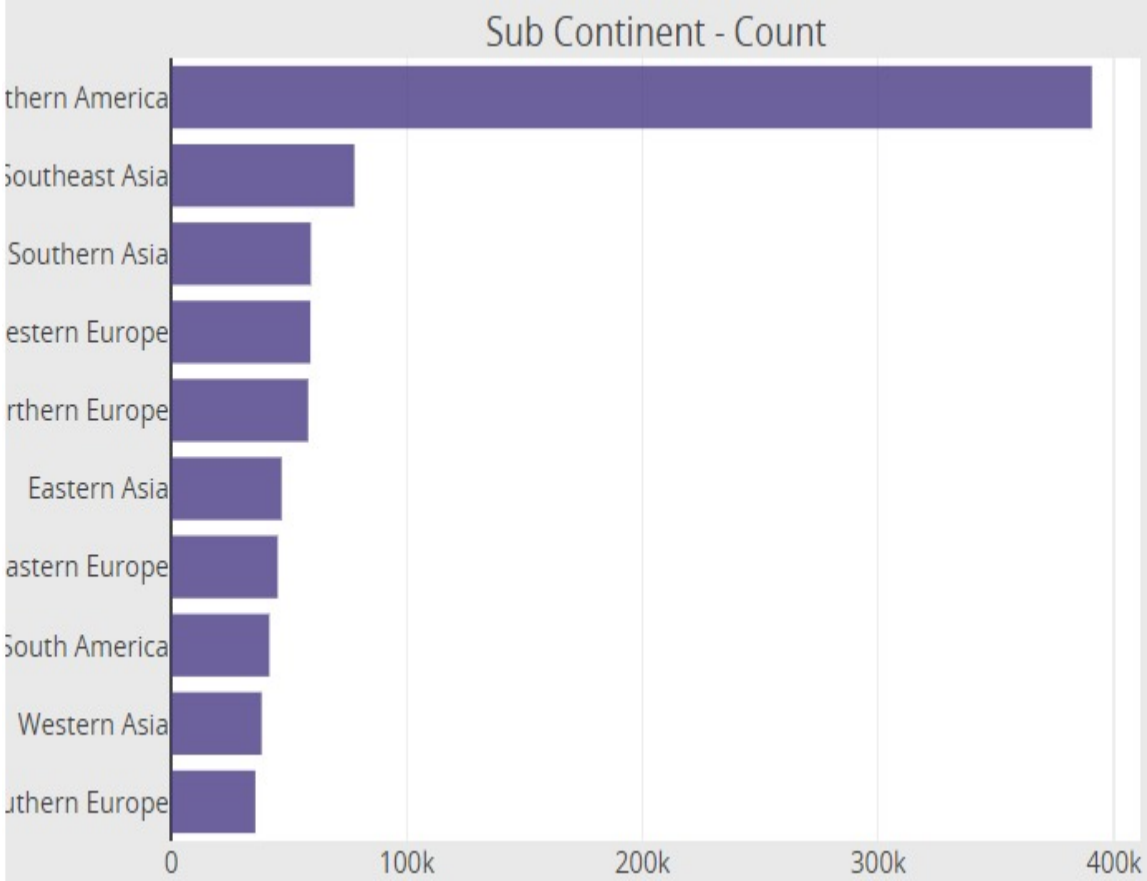
Preprocessing Analysis





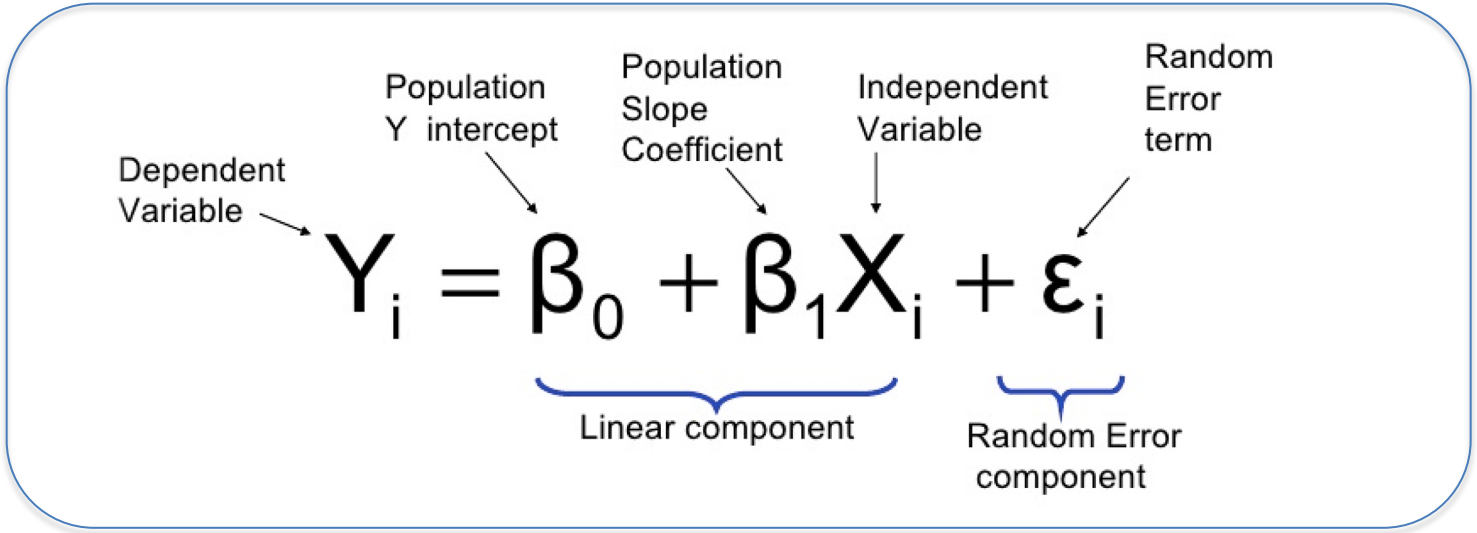
Preprocessing Analysis





Linear Regression Algorithm

- Linear Regression is arguably the most commonly used algorithm.
- Simple
- Less computational memory



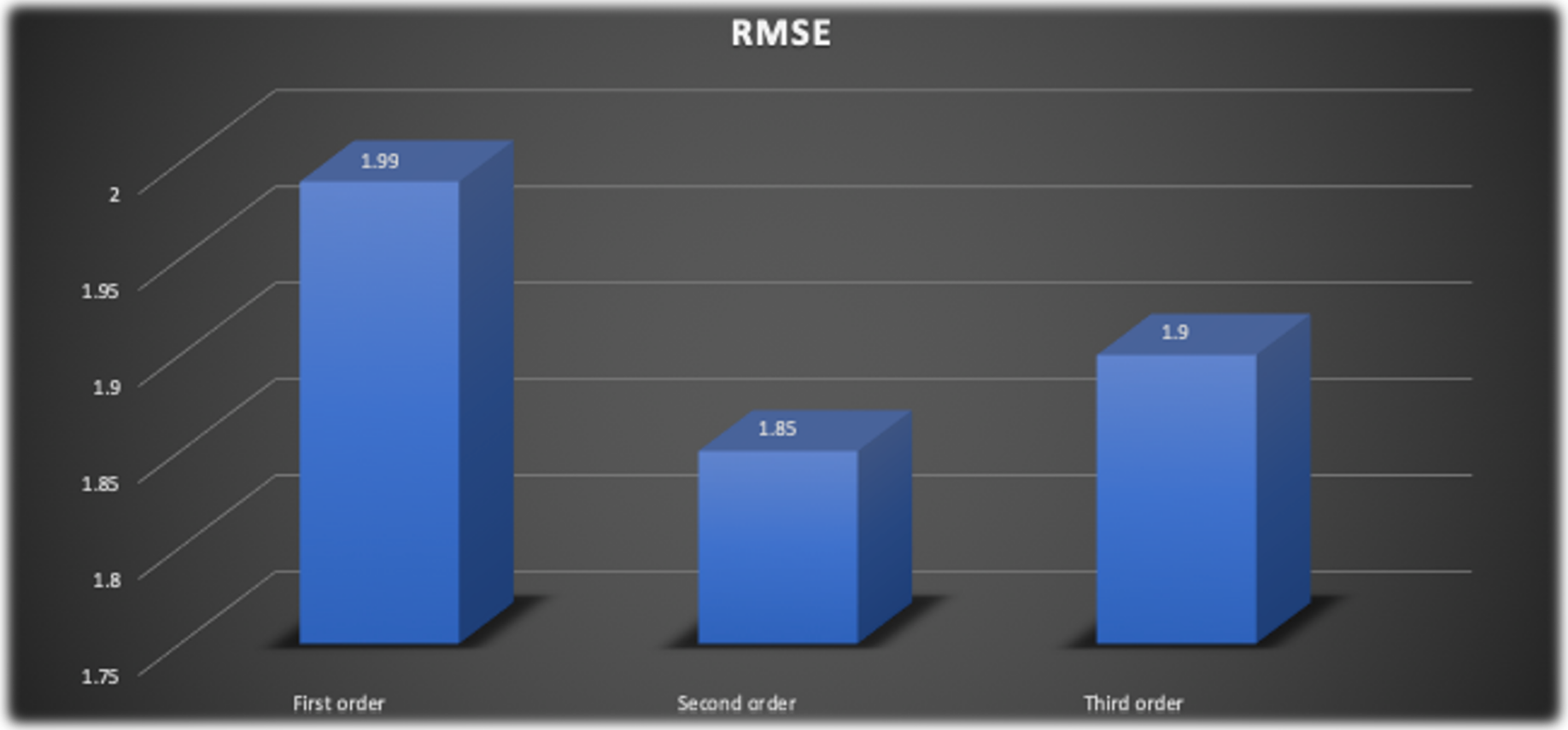
Error Analysis: (Root Mean Square Error)

Submissions are scored on the root mean squared error. RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where y hat is the natural log of the predicted revenue for a customer and y is the natural log of the actual summed revenue value plus one.

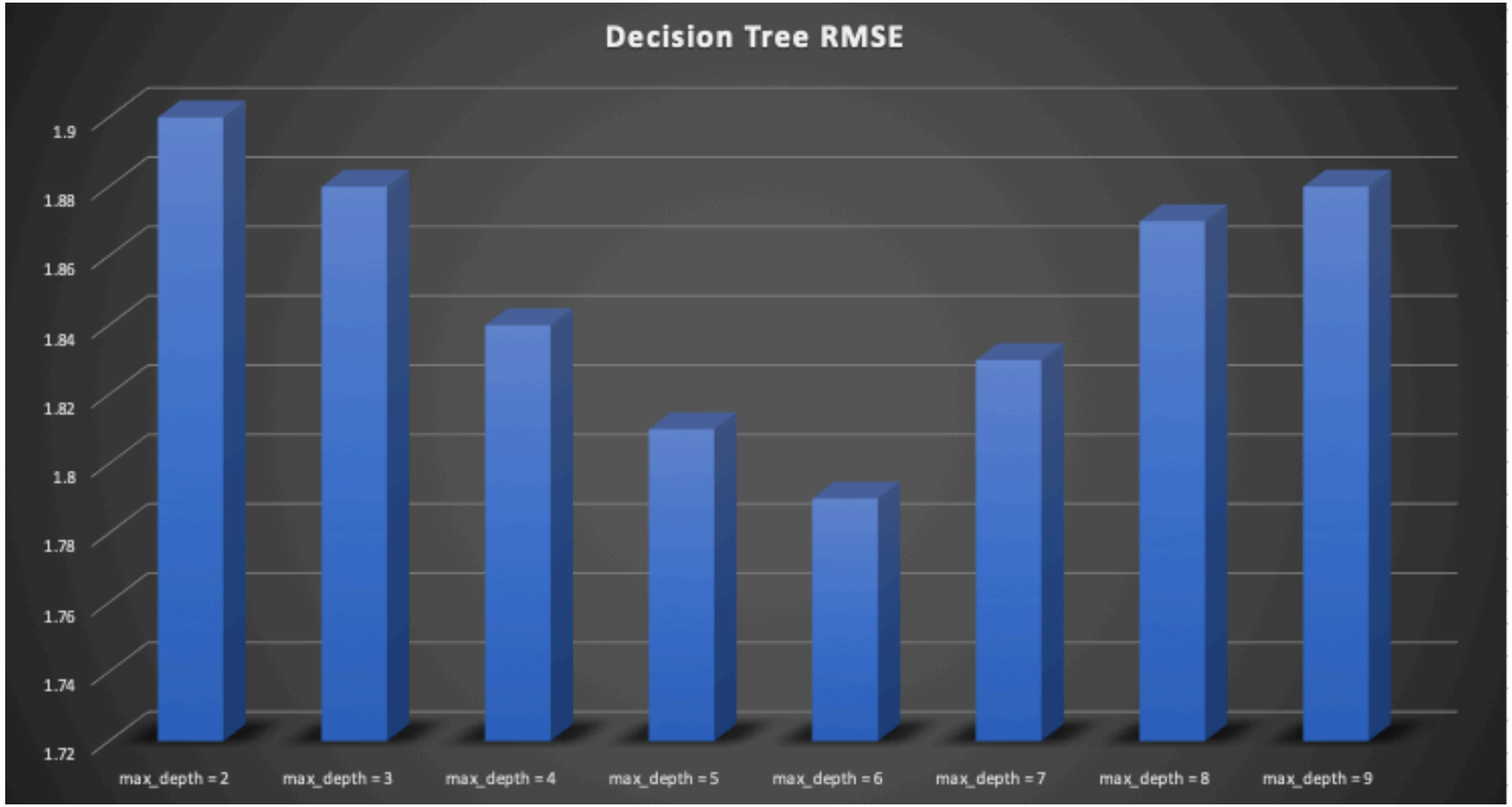
RMSE - 1st Order, 2nd order & 3rd order of Linear Regression



Algorithm – 2 Decision Tree Algorithm

- Nonlinear relationships between parameters do not affect tree performance
- Easy to understand and interpret.
- Easy to handle missing values without needing to resort to imputation.
- Can handle both numerical and categorical data.
- Perform well on large datasets
- Are extremely fast

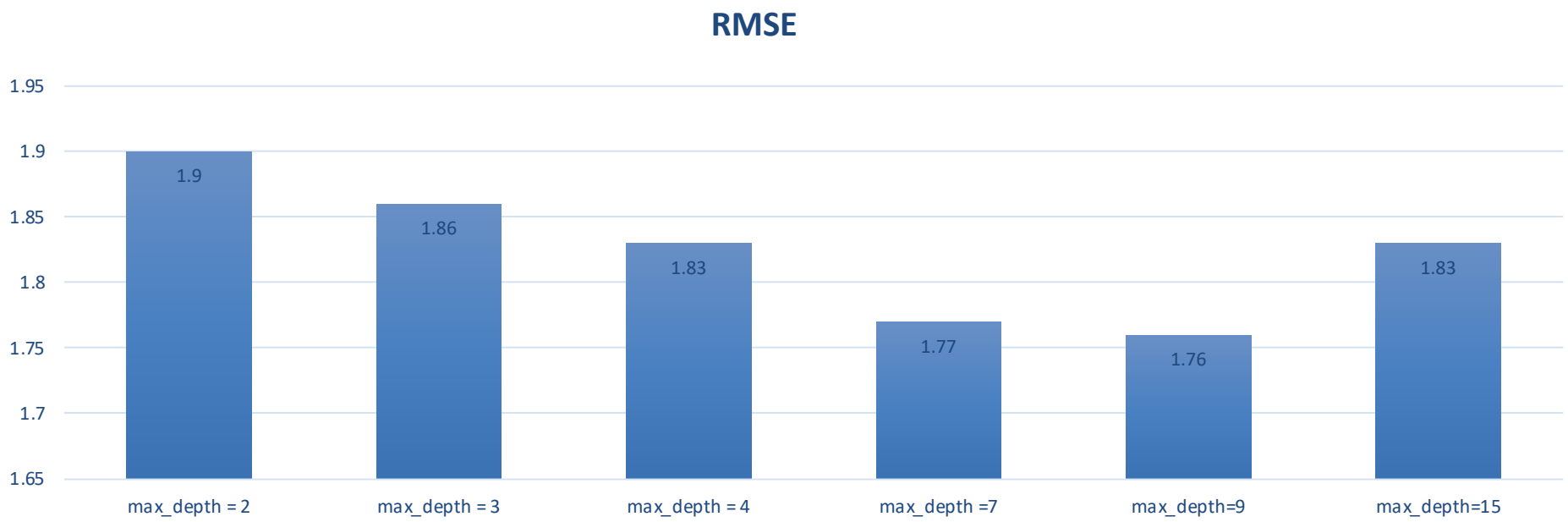
RMSE – Decision Tree (with different max_depth)



Random Forest Algorithm:

- Bootstrap Model
- As decision tree alone could not increase accuracy, moved to RF.
- Parameters used:
  - Max_length
  - Min_samples_split
  - Min_samples_leaf

Random Forest results based on Max_depth



Gradient Boosting Algorithm:

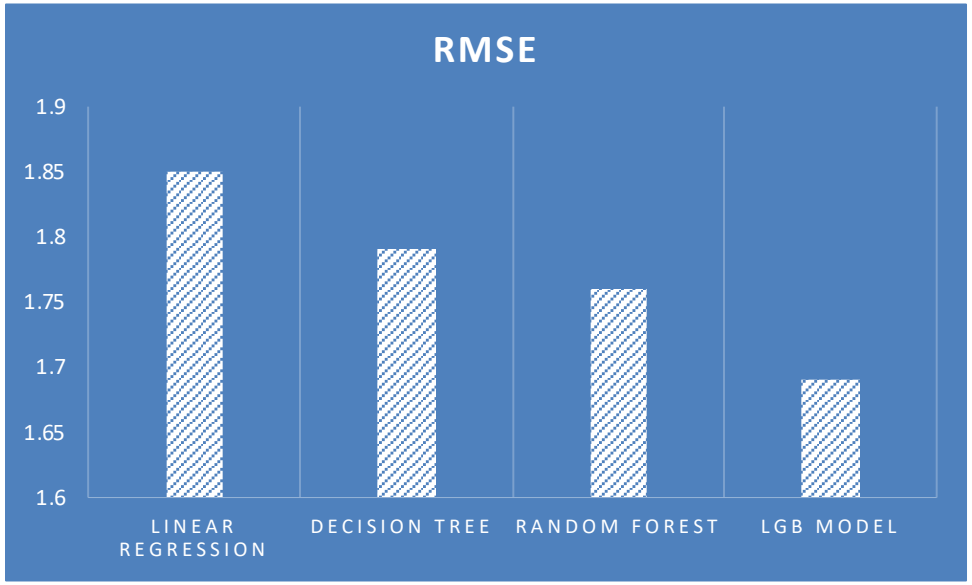
- LightGBM: A fast, distributed, high performance gradient boosting framework based on decision tree algorithms.
  - Microsoft Product
  - Bagging
  - Faster training speed and higher efficiency.
  - Better accuracy.
  - Capable of handling large-scale data.
  - Algorithm is built on decision trees. Aggregating models based on error of previous model.

Parameters used in LGB

- Num_leaves: 30
- Min_child_samples:100
- Learning rate:0.1
- Bagging fraction: 0.7
- Bagging frequency:0.5

RMSE error comparision for four models

Models	RMSE (Root Mean Square Error)
Linear Regression	1.85
Decision Tree	1.79
Random Forest	1.76
LGB Model	1.69



Work Distribution:

Aditi:	Data Pre-processing and Linear Regression Implementation
Ashna:	Data Preprocessing and Data Analysis
Koushik:	Data Preprocessing, Decision tree implementation
Madhusudhan:	RandomForest and Gradient Boosting Algorithm Implementation

What we have already employed

- Linear Regression
- Decision Trees
- Random Forest
- Gradient Boosting

What we look forward to use

- Model tuning
- Variable transformation
- Mostly your suggestions

SJSU SAN JOSÉ STATE UNIVERSITY