Revenue Forecasting for E Commerce Businesses

Utkarsh

Amit

Aadarsh

Department of ECE, IIITD

Abstract

Analyzing revenue data is essential for understanding the performance of a company's products and overall business strategy. By analyzing revenue data, businesses can identify top-performing products, monitor sales trends, forecast future revenue, and identify underperforming products that may require adjustments to marketing strategies or product development. This information is valuable for making informed decisions about resource allocation, inventory management, and overall business planning. We will be computing each of our expectations with 3 machine learning strategies at present which are linear regression, neural network, and decision tree. This multitude of strategies will be used to fit every one of the ideal expectations. The models will be tried with numerous hyperparameters and their exhibition will be examined.

Keywords: Profit and Revenue Prediction, Machine Learning, Decision Tree, Neural Networks, Linear Regression.

1. INTRODUCTION

1.1 Background

- What is E-commerce, and why do we need it? Why an ML model for forecasting revenue model ?

  E-commerce, short for electronic commerce, alludes to the trading of labor and products through web-based channels, for example, sites, versatile applications, or online entertainment stages. With the fast development of the web and innovative headways, online business has turned into an essential piece of present-day business. An ML model for estimating income can give organizations exact and dependable expectations of future income in view of verifiable information. ML models can give a more complex and computerized way to deal with income determining, permitting organizations to pursue information-driven choices and streamline their exhibition.

1.2 Literature Survey

- Revenue forecasting is a critical aspect of managing e-commerce businesses, and numerous studies have explored various approaches to predicting revenue for online retailers. The following are some of the surveys:
- One popular approach to revenue forecasting is the use of time series analysis, which involves analyzing past revenue data to identify patterns and trends and predict future revenue. (Wei, Peng, Yig, 2014)[3]
- Another approach to revenue forecasting is the use of customer behavior data. By analyzing customer behavior, such as purchase history, browsing patterns, and social media activity, businesses can predict future revenue and tailor their marketing strategies accordingly (Patangia, 2020)[2]

1.3 Objectives

Following are the objectives which we are aiming to achieve:

- Performing EDA on various features of the dataset.
- Net revenue and profit of the company for a given year.
- Calculating profit and revenues of following categories of the company with help of different ML models
    1. Country
    2. Sub-category feature
    3. Age - group

1.4 Scope

The scope of revenue forecasting for e-commerce business organizations utilizing ML is immense. ML models can be utilized to figure income for individual items, and deals channels, like web-based commercial centers, online entertainment stages, and retail locations. This can assist organizations with enhancing their item contributions, valuing techniques, and deals and appropriation methodologies and allot assets actually.

1.5 Impact

ML models are equipped for breaking down tremendous measures of information and distinguishing designs that traditional statistical models may miss. This can prompt more precise income figures, assisting organizations with pursuing better-informed choices, which assists companies with enhancing their stock levels, diminishing the risk of stockouts and an overabundance of stock and hence saving expenses.

2. MATERIALS AND METHOD

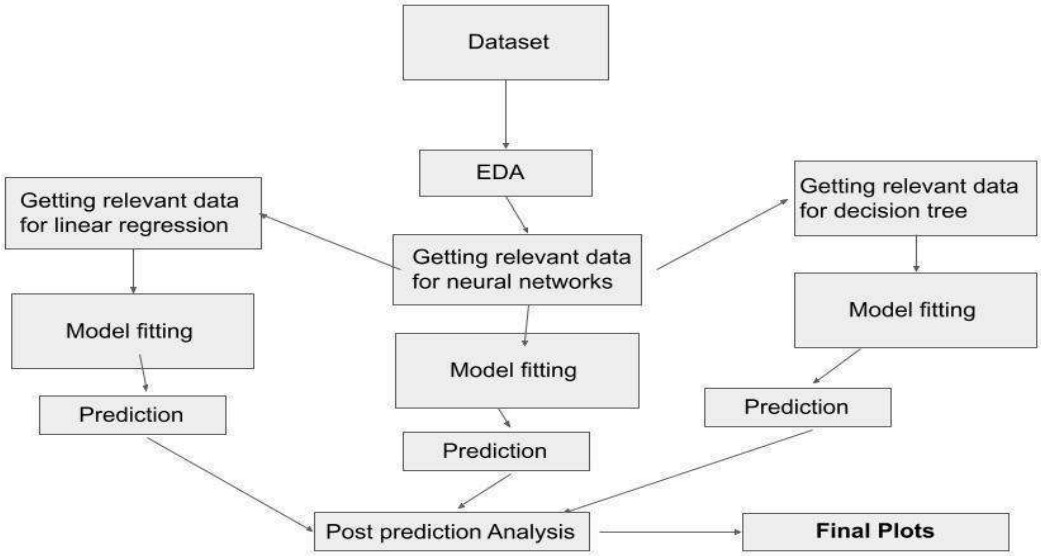
2.1 Dataset

We will be using a dataset of a company which has reach in multiple countries and sells a good variety of products. We have around 18 columns in our dataset in which profit and revenue are going to be our target vectors. We have features like date of purchase, category, sub category of products, customer age group, quantity purchased , unit cost, country of purchase , sex of customer etc. It has over 1 lakh 10 thousand plus data points which provides enough data for our model to get trained and also filtering relevant data out for varied prediction types. One can find the link to our dataset in the reference section.

Link: https://drive.google.com/file/d/1ExtvHACrwaiZ-AxxOz24EBIHN_uxkl_4/view?usp=share_l

	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Day	Month	Year	Customer	Age_Group	Customer	Country	State	Product_Categor	Sub_Categor	Product	Order_Qu	Unit_Cost	Unit_Price	Profit	Cost	Revenue
2	26	Novembe	2013	19	Youth (<25)	M	Canada	British Columbia	Accessories	Bike Racks	Hitch Rack	8	45	120	590	360	950
3	26	Novembe	2015	19	Youth (<25)	M	Canada	British Columbia	Accessories	Bike Racks	Hitch Rack	8	45	120	590	360	950
4	23	March	2014	49	Adults (35-64	M	Australia	New South Wale	Accessories	Bike Racks	Hitch Rack	23	45	120	1366	1035	2401
5	23	March	2016	49	Adults (35-64	M	Australia	New South Wale	Accessories	Bike Racks	Hitch Rack	20	45	120	1188	900	2088
6	15	May	2014	47	Adults (35-64	F	Australia	New South Wale	Accessories	Bike Racks	Hitch Rack	4	45	120	238	180	418
7	15	May	2016	47	Adults (35-64	F	Australia	New South Wale	Accessories	Bike Racks	Hitch Rack	5	45	120	297	225	522
8	22	May	2014	47	Adults (35-64	F	Australia	Victoria	Accessories	Bike Racks	Hitch Rack	4	45	120	199	180	379
9	22	May	2016	47	Adults (35-64	F	Australia	Victoria	Accessories	Bike Racks	Hitch Rack	2	45	120	100	90	190
10	22	February	2014	35	Adults (35-64	M	Australia	Victoria	Accessories	Bike Racks	Hitch Rack	22	45	120	1096	990	2086
11	22	February	2016	35	Adults (35-64	M	Australia	Victoria	Accessories	Bike Racks	Hitch Rack	21	45	120	1046	945	1991
12	30	July	2013	32	Young Adults	F	Australia	Victoria	Accessories	Bike Racks	Hitch Rack	8	45	120	398	360	758
13	30	July	2015	32	Young Adults	F	Australia	Victoria	Accessories	Bike Racks	Hitch Rack	8	45	120	398	360	758
14	15	July	2013	34	Young Adults	M	Australia	Victoria	Accessories	Bike Racks	Hitch Rack	7	45	120	349	315	664
15	15	July	2015	34	Young Adults	M	Australia	Victoria	Accessories	Bike Racks	Hitch Rack	7	45	120	349	315	664
16	2	August	2013	29	Young Adults	M	Canada	British Columbia	Accessories	Bike Racks	Hitch Rack	5	45	120	369	225	594
17	2	August	2015	29	Young Adults	M	Canada	British Columbia	Accessories	Bike Racks	Hitch Rack	7	45	120	517	315	832
18	2	Septembe	2013	29	Young Adults	M	Canada	British Columbia	Accessories	Bike Racks	Hitch Rack	2	45	120	148	90	238
19	2	Septembe	2015	29	Young Adults	M	Canada	British Columbia	Accessories	Bike Racks	Hitch Rack	1	45	120	74	45	119
20	22	January	2014	29	Young Adults	M	Canada	British Columbia	Accessories	Bike Racks	Hitch Rack	1	45	120	74	45	119
21	22	January	2016	29	Young Adults	M	Canada	British Columbia	Accessories	Bike Racks	Hitch Rack	1	45	120	74	45	119

2.2 Methodology



2.2.1 Exploratory data analysis:

We performed the following analysis on our dataset:

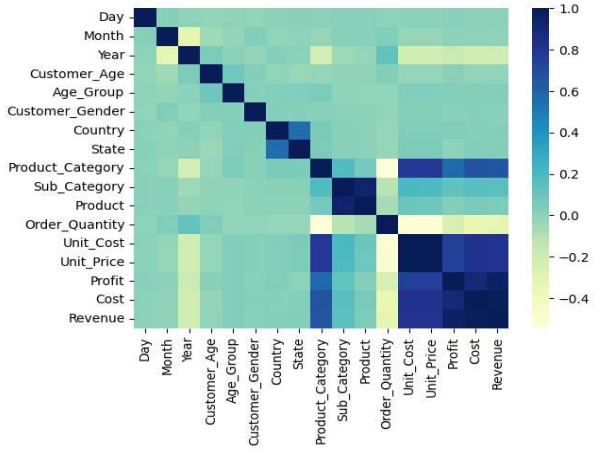
- Mean-Mode Standard Deviation, Minimum and Maximum value of each feature & Checking whether the null values are present or not:

	Day	Month	Year	Customer_Age	Age_Group	Customer_Gender	Country	State	Product_Category	Sub_Category	Product	Order_Quantity	Unit_Cost	Unit_Price	Profit	Cost	Revenue
count	113036.000000	113036.000000	113036.000000	113036.000000	113036.000000	113036.000000	113036.000000	113036.00000	113036.000000	113036.000000	113036.000000	113036.000000	113036.000000	113036.000000	113036.000000	113036.000000	113036.000000
mean	15.665753	6.453024	2014.401739	35.919212	1.197158	0.484129	2.191638	7.05193	0.609523	9.342970	50.156446	11.901660	267.296366	452.938427	285.051665	469.318695	754.370360
std	8.781567	3.478198	1.272510	11.021936	0.697601	0.499750	1.499522	6.53933	0.835298	4.516857	41.517561	9.561857	549.835483	922.071219	453.887443	884.866118	1309.094674
min	1.000000	1.000000	2011.000000	17.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	1.000000	1.000000	2.000000	-30.000000	1.000000	2.000000
25%	8.000000	4.000000	2013.000000	28.000000	1.000000	0.000000	1.000000	2.00000	0.000000	6.000000	9.000000	2.000000	2.000000	5.000000	29.000000	28.000000	63.000000
50%	16.000000	6.000000	2014.000000	35.000000	1.000000	0.000000	2.000000	4.00000	0.000000	10.000000	39.000000	10.000000	9.000000	24.000000	101.000000	108.000000	223.000000
75%	23.000000	10.000000	2016.000000	43.000000	2.000000	1.000000	3.000000	11.00000	1.000000	14.000000	95.000000	20.000000	42.000000	70.000000	358.000000	432.000000	800.000000
max	31.000000	12.000000	2016.000000	87.000000	3.000000	1.000000	5.000000	52.00000	2.000000	16.000000	129.000000	32.000000	2171.000000	3578.000000	15096.000000	42978.000000	58074.000000

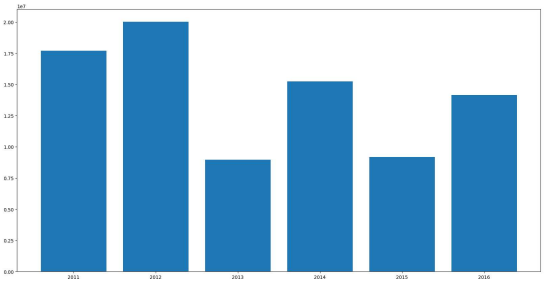
- Checking whether the null values are present or not & finding the correlation matrix:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113036 entries, 0 to 113035
Data columns (total 18 columns):
 #   Column            Non-Null Count   Dtype
---  ------            --------------   -----
 0   Date              113036 non-null  object
 1   Day               113036 non-null  int64
 2   Month             113036 non-null  int64
 3   Year              113036 non-null  int64
 4   Customer_Age      113036 non-null  int64
 5   Age_Group         113036 non-null  int64
 6   Customer_Gender   113036 non-null  int64
 7   Country           113036 non-null  float64
 8   State             113036 non-null  float64
 9   Product_Category  113036 non-null  float64
 10  Sub_Category      113036 non-null  float64
 11  Product           113036 non-null  float64
 12  Order_Quantity    113036 non-null  int64
 13  Unit_Cost         113036 non-null  int64
 14  Unit_Price        113036 non-null  int64
 15  Profit            113036 non-null  int64
 16  Cost              113036 non-null  int64
 17  Revenue           113036 non-null  int64
dtypes: float64(5), int64(12), object(1)
memory usage: 15.5+ MB
```

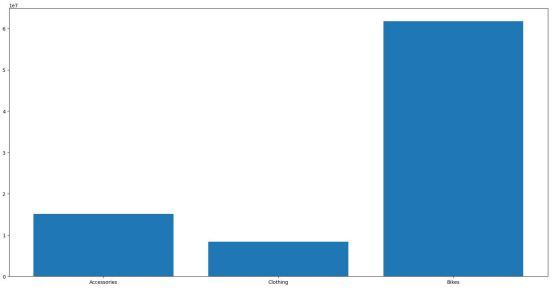
:



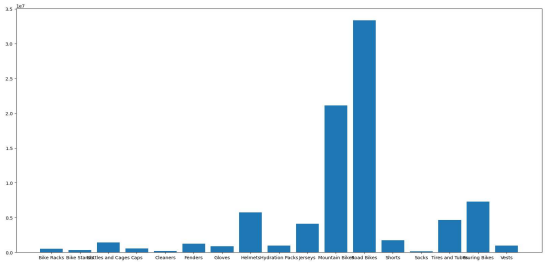
- Distribution Graphs
- Year v/s Revenue



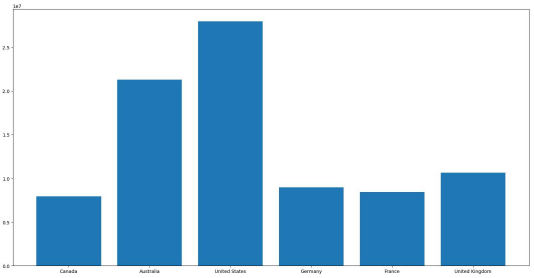
Product v/s Revenue



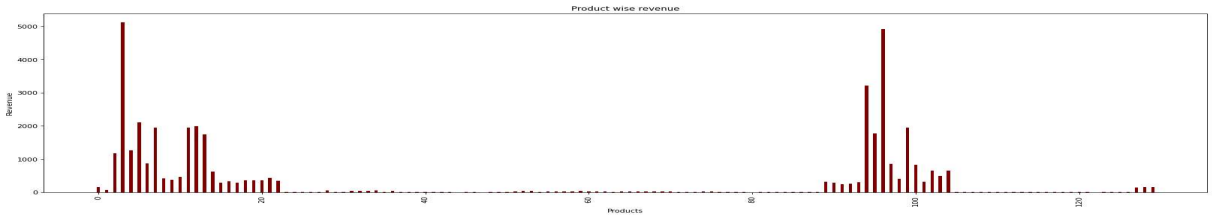
- Sub-Category v/s Revenue



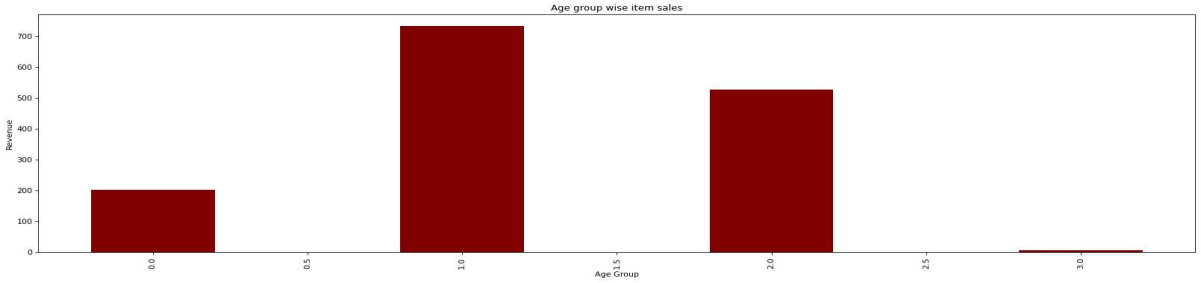
Country v/s Revenue



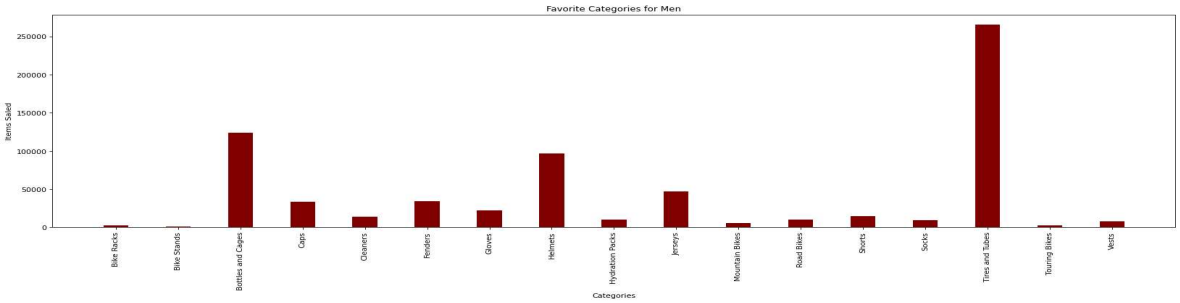
- Product v/s Revenue



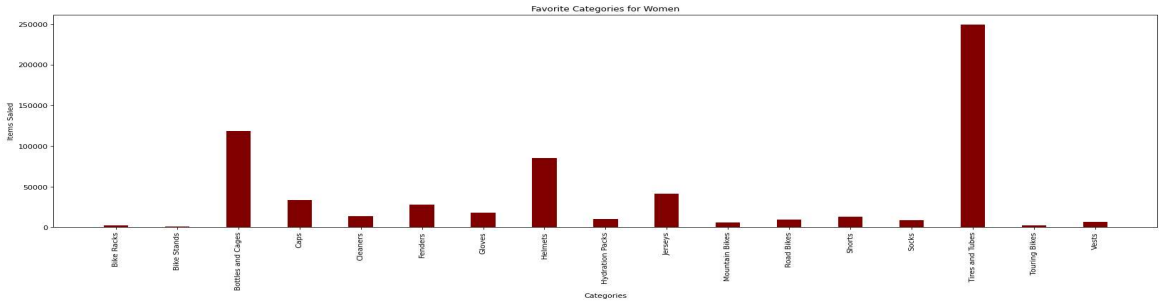
- Age Group v/s Item Sales



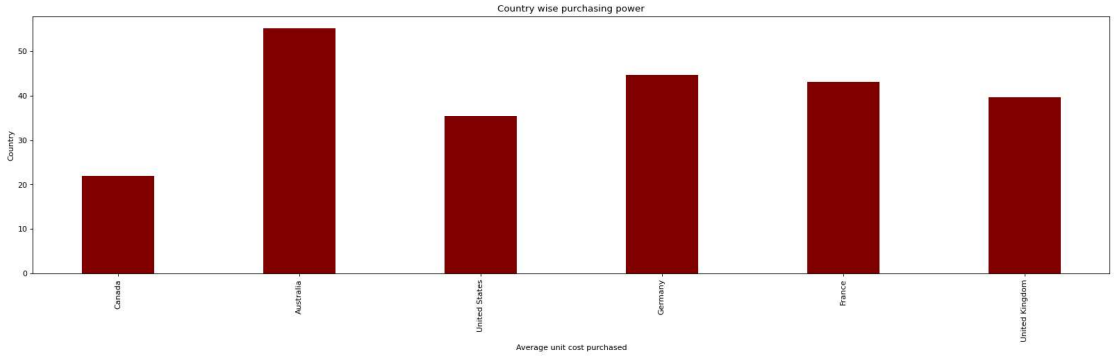
- Favorite Category for Men:



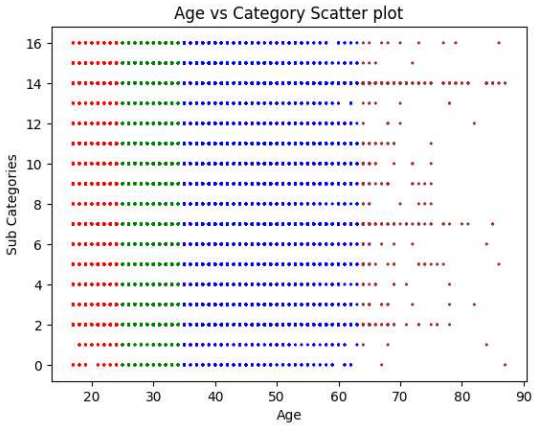
- Favorite Category for Women:



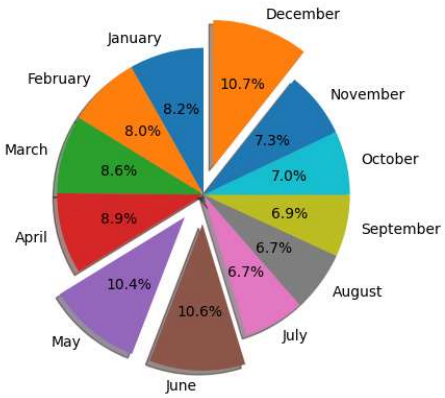
- Country- wise Purchasing Power:



- Age v/s Category Scatter Plot:



Month-wise Sale of Dataset:



- **Analysis after performing the EDA:**

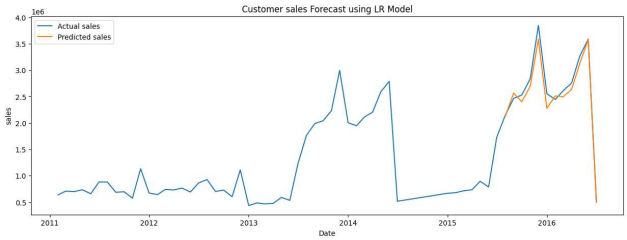
1. In order to normalize our dataset we have used the min-max scaler.
2. There were no null values found.
3. After finding the correlation matrix, we removed the highly correlated features.
4. Gender is not an important feature as there was no significant difference found in sales.
5. Bike category has the highest revenue.
6. The highest sales were noticed in December, June and May.
7. Most of the sub categories sales drop for customers aging > 65.

3. Model Selection, Training and Interpretation:

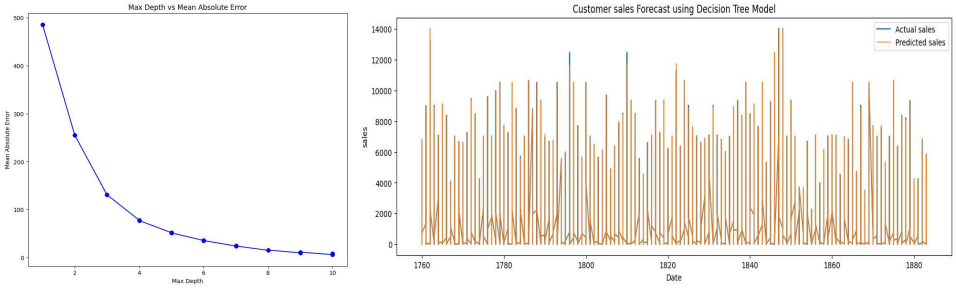
We have used the following models on our dataset:

- **Linear Regression:** The data was splitted into training(70%) and testing(30%). After removing the redundant features from our dataset, 10 of them were left. Thereafter we applied the linear regression model. Here are the observations:

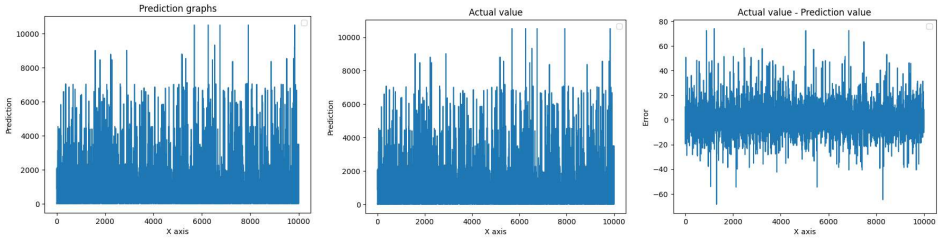
  RMSE: 56.0017613663



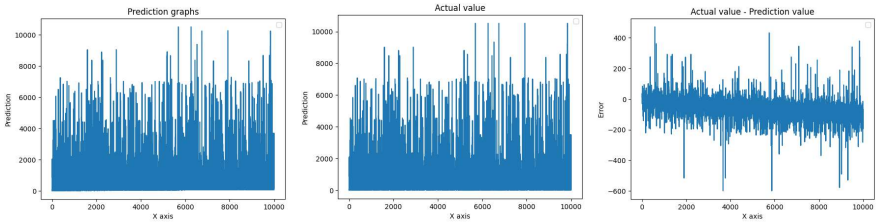
- **Decision Tree:** We have used the decision tree regressor here to explore for better results compared to the previous linear regression model. Here we are first dividing our dataset into train(60%), validation(20%) and test(20%) sets.
  RMSE: 17.552048499545403
  Now, we perform the **Grid Search** using **GridSearchCV** module to find the best hyper parameters for the regressor which we found out to be :
  max_depth : 10
  min_samples_leaf : 1
  min_samples_split : 2



- **Random Forest with K fold cross-validation:** We are trying to find a better RMSE but after applying K fold cross-validation our RMSE gets increased, where the value of k is 5.
  Average RMSE: 44.31867454822607
- **XGBoost:**
  Train RMSE error of best model: 8.230586751424866
  Test RMSE error of best model: 8.267245545611074

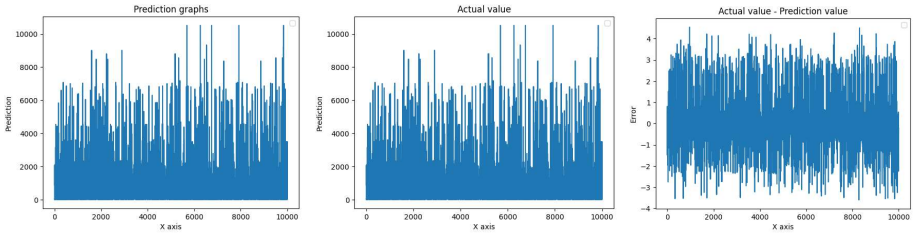


- **XGBoost Regressor with PCA:**
  Train RMSE error of best model: 25.57783189802092
  Test RMSE error of best model: 76.49888789902742

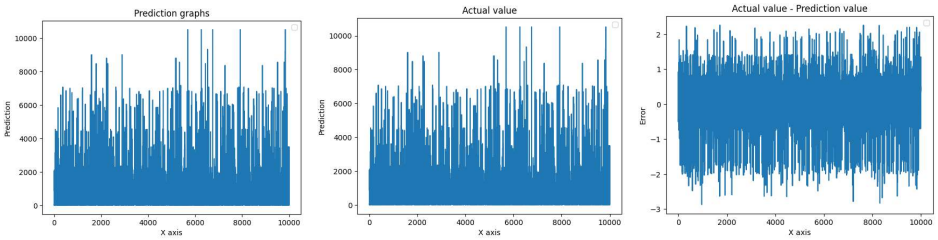


We have used **Grid Search** using the **GridSearchCV** module to find the best hyper parameters. Based on the outcome, the XGBoost regressor without PCA has performed significantly better than the XGBoost regressor with PCA. One of the reasons is that the PCA transformation is not preserving enough information from the original features, leading to a loss of predictive power.

- **Elastic Net**:
  Train RMSE error of best model: 1.021254564952723
  Test RMSE error of best model: 0.8255877717428844



- **Elastic Net with PCA:**
  Train RMSE error of best model: 0.7208423018024691
  Test RMSE error of best model: 0.6139800567406719



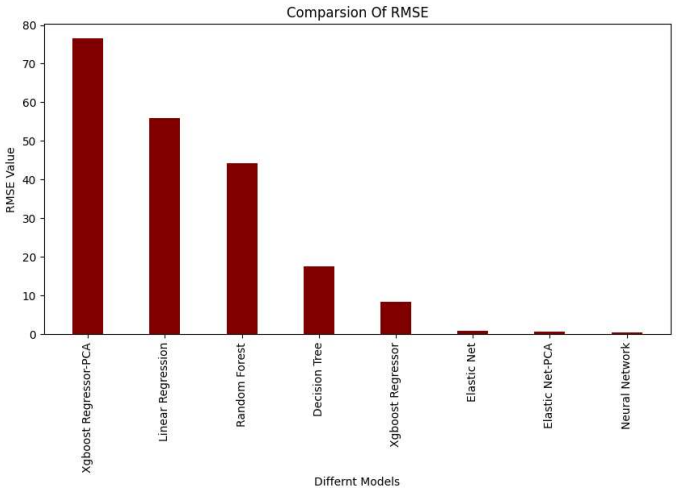
  As we saw earlier, the LR model was overfitting our data which led to its bad performance. Hence, to reduce the overfitting issue we have used Elastic Net regularization technique which is a combination of both L1 and L2 regularization. As showcased in the results of every model, Elastic Net with PCA has given the best results for our dataset.

- **Neural Networks:**
  RMSE Test: 0.3377558452724246
  RMSE Train: 0.3171605123805297
  In the neural network model there is one hidden layer with 32 neurons and a ReLU activation function, and an output layer with one neuron and a linear activation function. The model is trained for 100 epochs with a batch size of 32 using the fit method of the model.

4. CODE: LINK

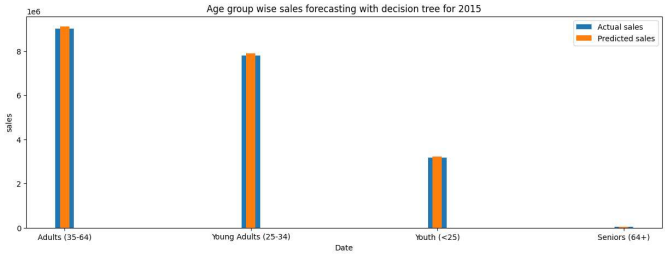
5. Inferences:
`

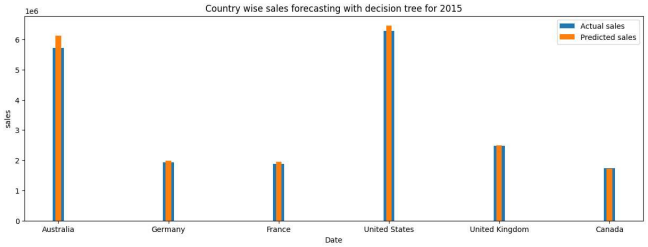
In our analysis we found out that linear regression was not able to perform really well with overfitting being one of the causes. Random forest is able to perform a little better with cross validation applied. Decision tree further improved the performance with the best hyperparameters found with the Grid Search. Further we also tested the XGBoost Regressor model for trying gradient boosting and found its best hyper parameters through grid search. Another drastic improvement was found with applying ElasticNet with PCA where dimensionality was reduced to 5 which reduced the complexity of the model and combination of L1 and L2 regularization techniques were applied to further prevent overfitting and to get a much better fit on test data. Finally the best performance was found using Neural Networks.

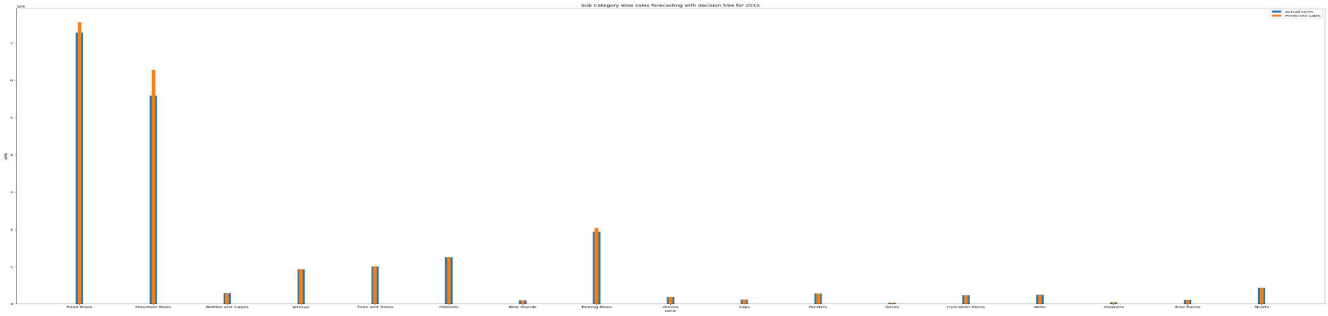


6. Novelty:

We have applied techniques like Grid search, gradient boosting and Principal Component Analysis to optimize our results which are not being worked upon much in the domain of sales forecasting. We have also added the category wise predictions below to show the versatility of our trained model.







7. References:

[1] Hsieh, P. H. (2019). A Study of Models for Forecasting E-Commerce Sales During a Price War in the Medical Product Industry. HCI in Business, Government and Organizations. ECommerce and Consumer Behavior, 3–21. https://doi.org/10.1007/978-3-030-22335-9_1

[2] Soham Patangia. (2020). Sales Prediction of Market using Machine Learning. International Journal of Engineering Research And, V9(09). https://doi.org/10.17577/ijertv9is090345

[3] Wei, D., Geng, P., Ying, L., & Shuaipeng, L. (2014, May). A prediction study on e-commerce sales based on structure time series model and web search data. The 26th Chinese Control and Decision Conference (2014 CCDC). https://doi.org/10.1109/ccdc.2014.6852219.