Suhail Sharafuddin

Amazon Sales Report For 2019

Problem Statement

Many companies are using predictive analytics to predict future sales. It is becoming more and more important for companies to be able to predict future sales. The advantages of doing this are plentiful. It quantifies growth, propels growth, and aids in target attainment.

I will be using Amazon sales data for 1 year from Kaggle. The data ranges from January 2019 to January 2020. Amazon is a company that highly benefits from predictive analytics, since there is a huge amount and variety of products being sold through it.

In this project, I will develop a machine learning model to accurately predict future sales for a given period based on historical sales data, to assist Amazon in making informed business decisions, optimizing inventory management, and enhancing overall sales performance.

Data Wrangling

First, I had to do the data wrangling. Data wrangling is an incredibly important step. It is the process of transforming the raw data into a more consumable format for machine learning. It involves making the data presentable and cleaning up errors.

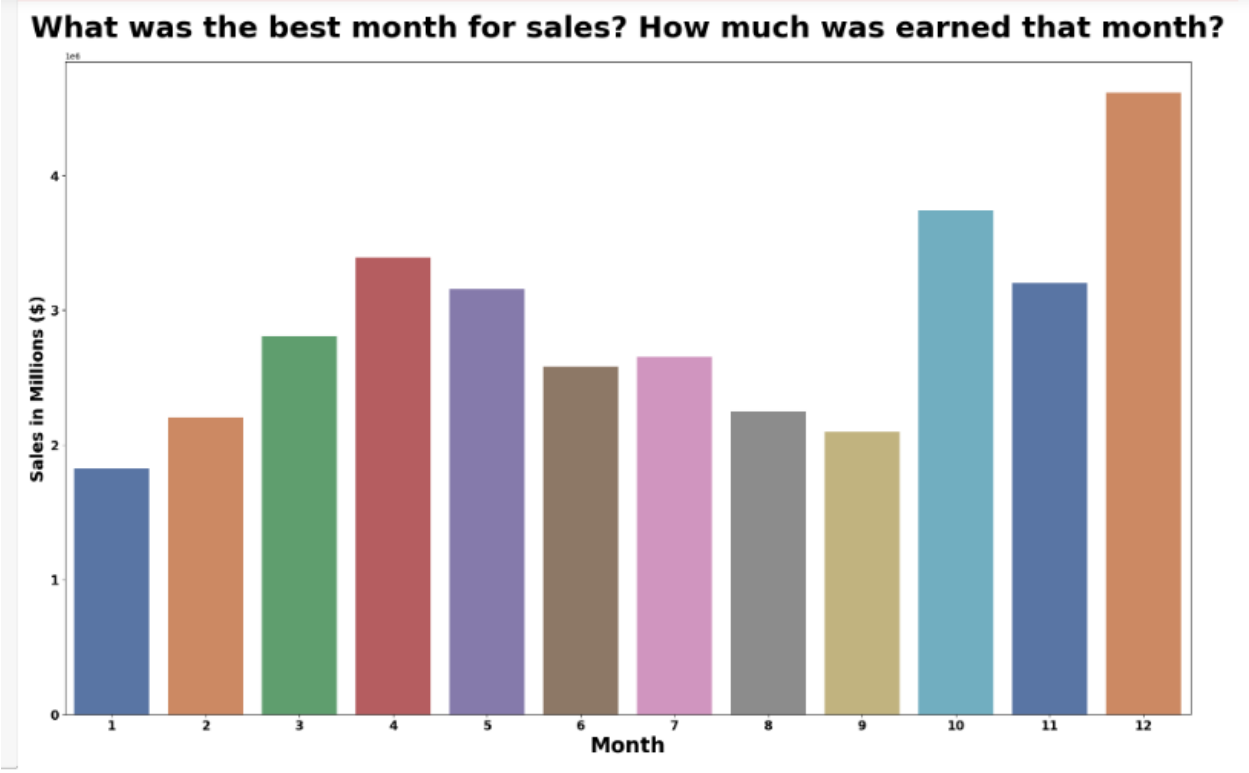
Since the dataset was split up into 12 different files, for each of the 12 months, I had to concatenate them into one big file. Then, I had it in order by date. It was different to having it be ordered by Order ID. I decided having it ordered by Order ID was the better choice. I realized any time there were NAN values, the entire row was filled with NAN. So, I dropped all NAN values.

Exploratory Data Analysis

Exploratory Data Analysis is getting a feel for your data, usually by visualization. It helps summarize the main characteristics of the dataset.

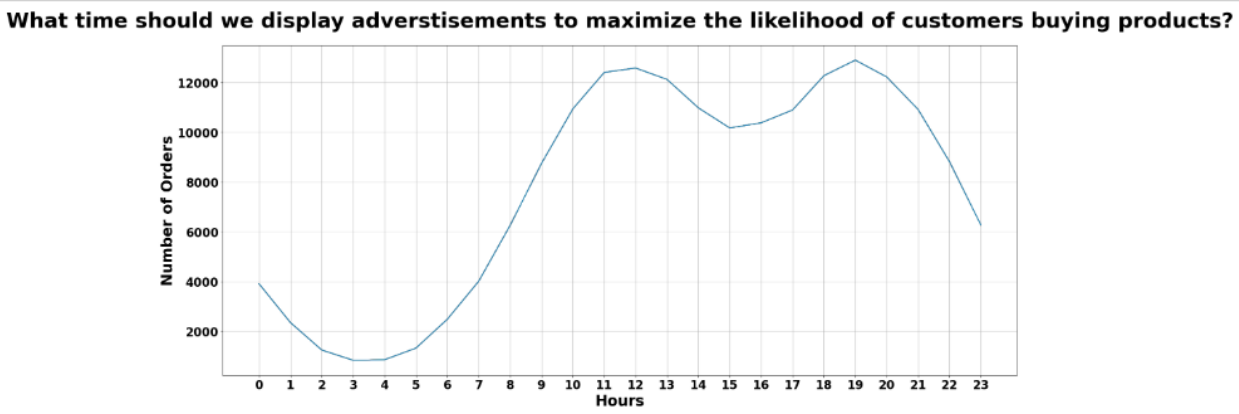
Before I dove into the analysis itself, I decided adding some columns would be helpful. First, I changed the Order Date column into a datetime format. Then, I extracted Year, Month, Hour, Minute values and put them into separate columns. Also, I created separate columns for Sales, and City. Sales would be the unit price multiplied by the quantity ordered by the user.

Then, I visualized the most important characteristics of the dataset. The first visualization is the month with the highest number of sales, and how much was earned for that month. The next visualization is the number of orders at different times of the day. This was to see what time would be best to display advertisements to maximize the likelihood of customers buying products.



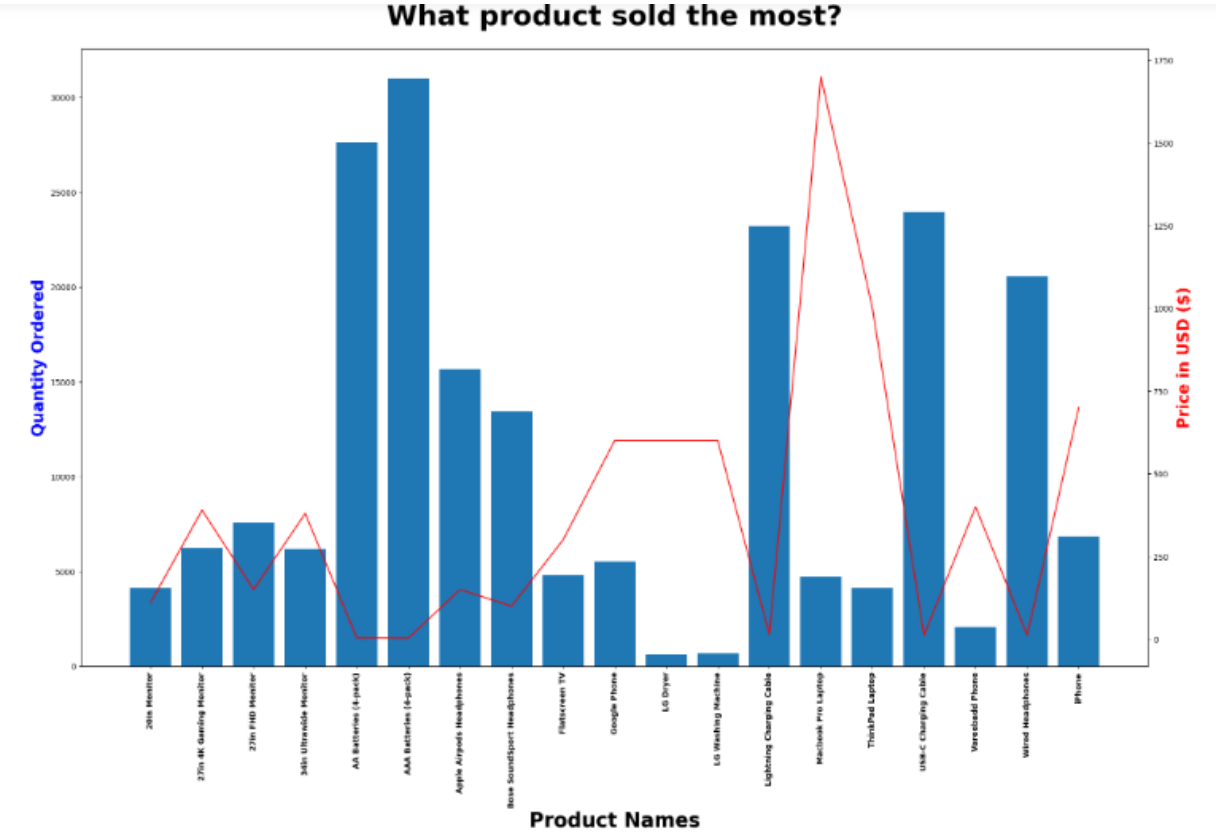
According to the graph, the best month to sell is December. The number of sales exceeds $4.5 million. This is most likely because of the holiday season that takes place in December.

The next visualization is the number of orders at different times of the day. This was to see what time would be best to display advertisements to maximize the likelihood of customers buying products.



The best times to display advertisements to maximize the likelihood of customers buying products is 9 AM to 10 AM, and 4 PM to 5 PM

Finally, the last visualization is to see the products that sold the best, based on both quantity and price.



The most sold products are AAA Batteries (4-pack), AA Batteries (4-pack), Lightning Charging Cable, USB-C Charging Cable, and Wired Headphones. The reason for this is most likely because these products have low unit prices, so more units were bought.

I also checked which products were most likely to be sold together. The two products that are sold together the most are iPhone and Lightning Charging Cable with 1005 orders, while the Google Phone and USB-C Charging Cable are the second highest with 987 orders.

```
('iPhone', 'Lightning Charging Cable') 1005
('Google Phone', 'USB-C Charging Cable') 987
('iPhone', 'Wired Headphones') 447
('Google Phone', 'Wired Headphones') 414
('Vareebadd Phone', 'USB-C Charging Cable') 361
('iPhone', 'Apple Airpods Headphones') 360
('Google Phone', 'Bose SoundSport Headphones') 220
('USB-C Charging Cable', 'Wired Headphones') 160
('Vareebadd Phone', 'Wired Headphones') 143
('Lightning Charging Cable', 'Wired Headphones') 92
```

Preprocessing

Now that my data was cleaned and I better understood the information that it held, I was ready to get started on building models. There were a few steps I had to take to prepare my dataset for modeling though.

I made sure to separate the Purchase Address into State and City columns. Then, I used label encoding for the Product, State, and City columns. I also created a column that would show the difference between the previous day's sales and the current day's sales called sales_diff. Finally, I split the data into testing and training datasets.

Model Selection

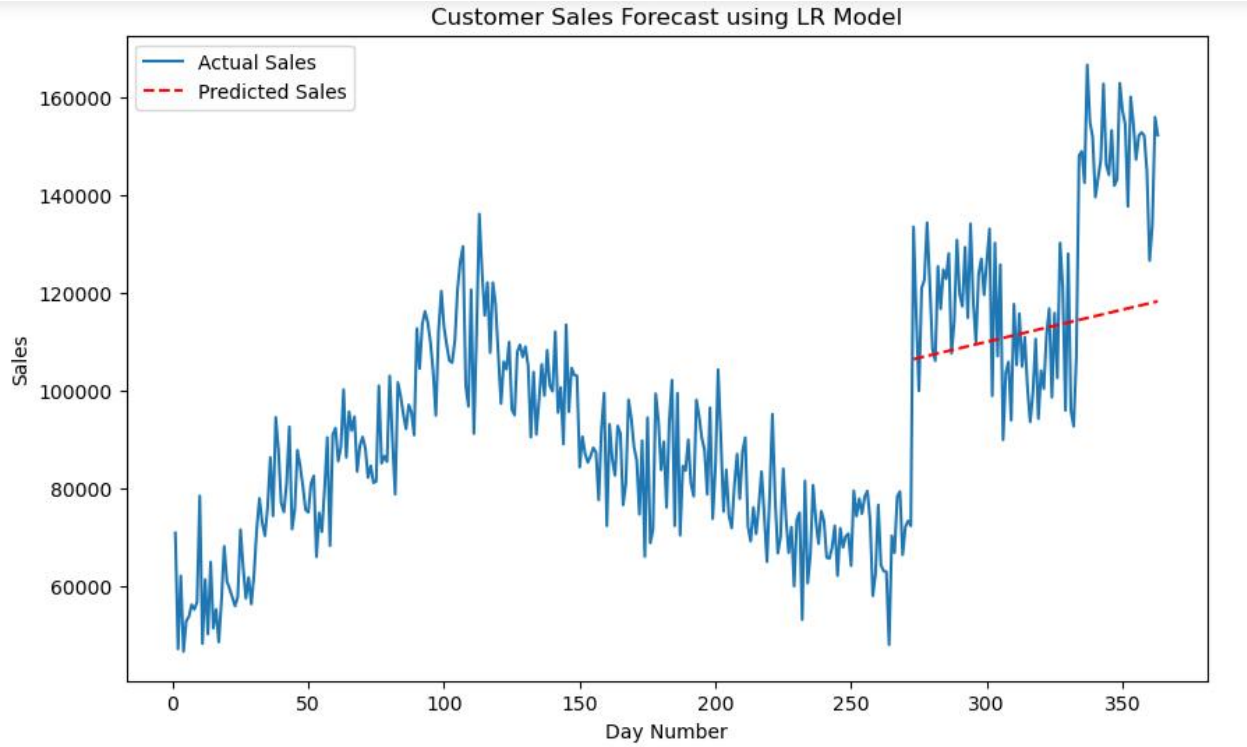
I was looking at which machine learning models would be the best for the dataset that I had. The metric that I focused on when building my models was accuracy. Specifically, I wanted to find the R^2 score.

I recognized that for this specific problem, Random Forest Regressor would be the best choice for my model. I decided to also use Linear Regression as well. Below shows which models would be the best for this:

Method
RandomForestRegressor
LinearRegression
Lasso
DecisionTreeRegressor
KNeighborsRegressor

Before building the models, I needed to do some feature selection. The features I chose for Linear Regression model are the Total Revenue, Price Each, Quantity Ordered, and the sales_diff column. For the Random Forest Regressor, I kept the Total Revenue, Week Number, Day Number, and WeekDay Number columns.

Linear Regression was not the best at giving accurate predictions. Here is the plot of the Linear Regression model:

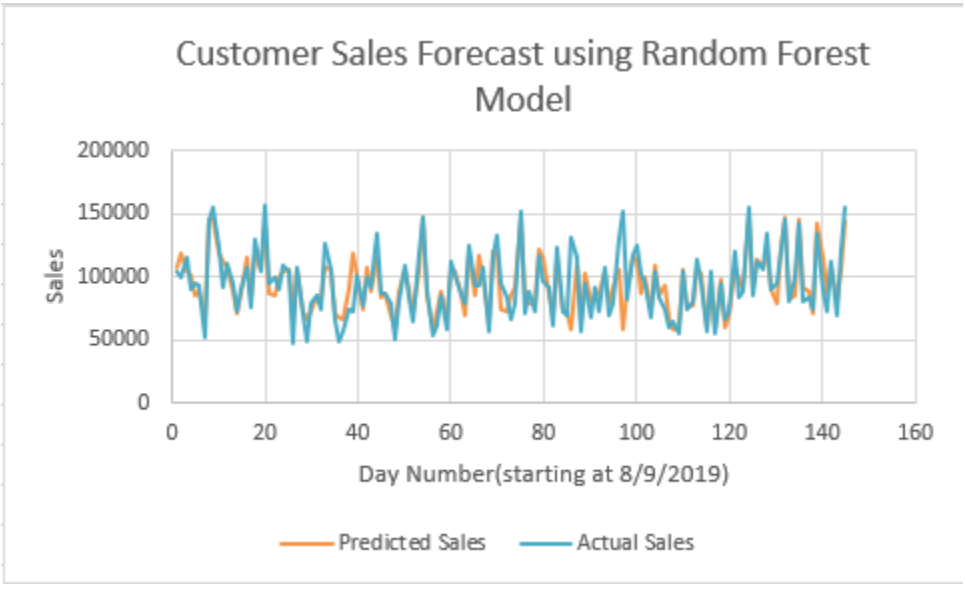


Here is the metrics of the Linear Regression model:

```
Linear Regression MSE:  126774.76768423527
Linear Regression MAE:  125158.0342857143
Linear Regression R2:  -23292523.23390119
```

The R^2 is the one metric that I really care about. Linear Regression was not the best option for this model since the data could not easily be explained by a line of best fit.

The Random Forest was a much better model. Here is the plot of the Random Forest:



Here is the metrics of the model:

```
Random Forest R2:  0.5823861087114155
```

When it comes to the models, Random Forest gave me a legitimate R^2 , unlike the Linear Regression model. Increasing the number of trees in the forest increased the R^2, but not by much. This is most likely the highest the value will be the data currently.

Takeaways

The Random Forest model was the best model for this project. It did a pretty good job with the data that it was given. However, the lack of data did not allow the R^2 to be any higher. Having only 1 year of data was not enough for the model to accurately predict at a higher rate than 58%.

Most of the features were not important for the models. There were redundancies in the parameters, and there were also parameters that ultimately had no impact on the target variable.

Future Research

Overall, this project gave me good experience with machine learning in a way that many companies would have experience with. Using Amazon and its wide variety of products as the dataset was helpful too.

Most importantly, I would like to expand the timeframe to include 3-5 years of data. The model would have more to work on, and the seasonality trends would be more pronounced. The accuracy would be much higher as well. If I want a higher accuracy score, 1 year of data would be not enough.

I would also like to try with other datasets from other companies' sales data. Amazon being the versatile e-commerce website meant that it was not as beholden to other factors, such as weather. Trying to run models where it would have an impact on the sales would be interesting.