Short Term Price Optimization For Amazon

Titiksha Amol Desai

Abstract / Introduction : The main goal of the project is to analyze the historical data for Amazon that has been scrapped from October 2023 and provide optimal prices for the products by using hybrid model, i.e., blend of Time Series model and Machine Learning models in order to achieve the dynamicity even though the data is historic.

Literature Review : E-Shopping has became a famous trend in today's world saving time of many people. Many such e-commerce platforms are available in the market theat provide these services. One of such yet famous e-commerce platform is Amazon. It is used by many buyers to do the online shopping. This site offers various services like quality products, introduction of new products everytime as well as well as fast and safe delivery. Like many other platforms (Uber for instance), Amazon has also adapted optimal pricing strategy that dynamically optimizes the prices of product. Amazon uses several Time Series, Regression, Machine Learning models, Neural Networks as well as Optimization Algorithms to achieve this goal. All these models and algorithms are used independently. Thus, providing them a robust solution for optimizing the pricing strategies. The analysis and modeling is done on the latest available historical data (say till yestaurday) updated on daily basis. Amazon passes this daily updated historical data to these independent models. Thus allowing them to obtain dynamic optimized price. In this way Amazon's present algorithms optimizes the price of product dynamically, helping them to improve sales everyday and then increase the profitability. My approach stands unique in terms of Amazon's existing approach is beacause firstly I have worked on short term historic data source which is scrapped dataset from October 2023 obtained from the public data source Kaggle. This helps my models to run fast due to limited data source. Secondly, I have used hybrid model approach that is blending Time Series model with the Machine Learning models which is unique approach if we compare to Amazon's existing approach. This hybrid model helped me achieve robustness in the model since it was combination of ARIMA and Random Forest as well as LGBM. I compared which model gives less RMSE. Furthermore, I hypertuned Random Forest as well as LGBM model and have added result from the model that gave me least RMSE. Using these results as well as other important features : ARIMA Forecasted Prices, Popularity Score, Market Share, Price Elasticity and Inventory Turnover Rate. Said this, the results have been achieved by using the finest modeling strategies.

Methodology : To start with, I have taken statastical historic dataset scrapped from October 2023 where the data resource is the famous public data resource Kaggle. For this project, I am using tool jupyter and language Python. At first I have loaded data using Pandas library then I have read it. Upon reading the data, I did first clean, that is, removed the columns that weren't required and renamed few columns as per convinience that made data reading easy. Second step I have performed is Feature Engineering. This was the crucial step in this project since the features required for the goal of the project were not present initially. By feature engineering, I generated those features in my dataset. Those are of 3 categories, first reflects the features required for observing price trend as this is what we are trying to optimize throughout the project. Second category reflects the consumer behaviour and market trend. This category helps in understanding consumer behaviour as well as market trend since these factors affects the pricing strategy and plays an important role in achieving the goal of project. Third category reflects the features required for time series model. Since the blend of time series with machine learning models adds the dynamicity prices or in simple words, blend of time series helps in optimizing the prices. After Feature Engineering, I performed second data clean meaningly removing null and NaN values from the dataset. After performing second data clean I performed Exploratory Data Analysis (EDA) where I first understood data using panda functions. Then I desingned some visulazitions that helped me understand current situation. These visualizations also fall in 3 categories. These three categories are Price, Consumer

Behaviour and Market Trend at the last but not least Time Series model. After analyzing these visualizations, I came to conclusion that optimal pricing strategies were needed to be implemented in order to improve the profibality for the given historical data. To achieve this goal, I first added date feature as this is the crucial feature required for developing time series model. Then I developed the ARIMA model. In order to develop hybrid model, I blended the output of time series with the machine learning models. The first ML model I have chosen is Random Forest Regressor. The output of Time Series model has been combined with the input of Random Forest Regressor model. Then I have tested the model using some metrics. The model has been hypertuned using GridSearch and tested again with the same above metrics. Significantly less RMSE has been achieved as compared to the without hypertuned model. Second ML model I have chosen is LGBM model where I followed similar process to make it hybrid model as I had followed for Random Forest Regressor model. Similarly, I tested the LGBM model with same metrics. Then I hypertuned this model and tested again with the similar metrics. A comparison report has been made for the four cases : Random Forest Regressor, LGBM Regressor model without hypertuning and Random Forest Regressor, LGBM Regressor with hypertuning. LGBM Regressor model with hypertuning has achieved less RMSE as compared to the other three cases. The output of hypertuned LGBM Regressor model has been used along with other important features : ARIMA Forecasted Prices, Popularity Score, Market Share, Price Elasticity and Inventory Turnover Rate in order to obtain the optimal prices. The last set of EDA (the visualizations that were done initially) has been performed in order to observe the difference if any. Optimal pricing strategy indeed improved conumer behaviour and market share thus increasing revenue, profitability, inventory turnover rate potentially.

To boost optimal prices I implemented following methods :

1. Upon comparing the four models : Random Forest Regressor, Hypertuned Random Forest Regressor, LGBM Regressor, Hypertuned LGBM Regressor the hypertuned LGBM Regressor performed best yielding least RMSE (429.95)
2. The predictions of hypertuned LGBM regressor were added into the updated dataset
3. Now, the max of ARIMA Forecasted Prices and the result of hypertuned LGBM Regressor were added to optimal prices
4. Next, for popularity score; if popularity score is greater than it's median it's results has been updated to the optimal price since popularity score influence high sales leading to profit
5. After this, for inventory turnover rate; if inventory turnover rate is greater than it's median it's results has been updated to the optimal price since inventory turnover rate too influence high sales leading to profit
6. Now the other key feature, if market share is greater than it's median it's results has been updated to the optimal price since market share influence high sales as well leading to profit
7. The last key feature, if price elasticity less than it's median, the results has been updated to the optimal price since lesser the price elasticity more the sales leading to increase in profit

Upon implementing these strategies by using the predictions of ARIMA model and hypertuned LGBM Regressor model along with other key features, I was able to boost the obtained optimal prices upon using the first strategy. These strategies indeed helped to improvise sales, revenue and profitability which has been evaluated by the growth of other important key metrics

Note Of The Changes Not Stated In Project Overview Statement:

Following changes have been employed when certain restrictions came into place while implementing the hybrid model

1. Downsampled the data to 40% : Due to blending the predictions of ARIMA model with other ML models, I faced a certain error : 'Pickling Error' or 'Task Unseriazable' which is basically the 'Memory Leakage' error. In order to resolve this error, I downsampled the data to 40%, earlier it was 100%. By downsampling data, we take only certain amount of data in order to avoid the memory leakage issues. Thus, by reducing the data sample to 40% of data, I was able to resolve memory leakage issue and all the hybrid models runned smoothly yielding the predictions and was able to go forward with other steps

2. Change Of ML Model : Earlier in my Project Overview Statement, I had chosen Gradient Boosting Machine (XGBM) as my second model after Random Forest Regressor model to implment. While implementing hybrid model, XGBM model took a very long runtime. When XGBM model as well as hypertuned XGBM models were implemented, though model ran but it took around 2 hours to run completely and yield the results. In order to reduce the runtime, I changed the model from Gradient Boosting Machine (XGBM) to Light Gradient Boosting Machine (LGBM) which is specifically designed for the large datasets. Upon implementing this model, the overall runtime was reduced from 2 hours to 1 hour 15 mins, saving 45 minutes of runtime which is potentially good improvement.

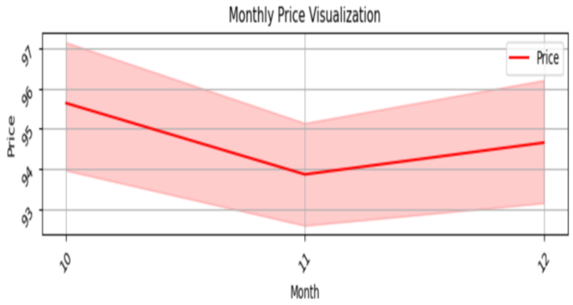
3. Additional Key Features To Boost Optimal Price : Earlier in Project Overview Statement, I had defined to use only the predictions of best performing hybrid model (Hypertuned LGBM Regressor Model) in this case as the Optimal Prices. But when I visualized Bullish/Berrish Signals with these optimal prices I saw the results were same. Seeing the same results even after developing robust models and using it's predictions as optimal prices, I decided to employ other key features as well in order to boost the obtained optimal prices. Upon employing the mentioned strategies in methodology above for boosting Optimal Prices, I was able to see the significant change in results. There was good improvement in sales, revenue and profitability which has been evaluated by the improvement in other important key metrics as well

Analysis : Upon performing basic data preparation and analyzing the given historic data by performing EDA, I first came to conclusion that there was drop of sales, reduced profitability, increased consumer dissatisfaction and overstockage of goods throughout the October to December that could due to various possible reasons but most likely reason could be ineffiecient optimal pricing strategies. Upon using the unique strategy for optimal pricing that is blending output of ARIMA model with the best performing model and efficiently using the results obtained alon with other key features like : ARIIMA Forecasted Prices, Popularity Score, Market Share, Price Elasticity and Inventory Turnover rate, I was able to obtain optimal prices thus improving the overall profitability for the given historic dataset.

Visuals : After performing the initial steps of any data science project, i.e., Data Loading, Data Cleaning, Data Preparation and Exploratory Data Analysis; following visuals depicts the requirement of optimal pricing for the gathered historical dataset.

(A) <u>VISUALS BEFORE DEVELOPING ML MODELS</u>

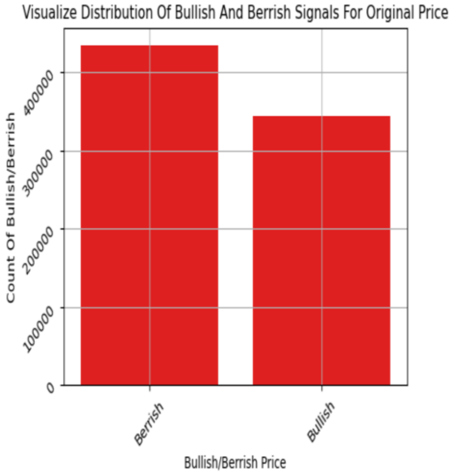
[1] Monthly Price Visualization



This visualization tells the drop of prices throughout the end of the year, thus impacting the overall sales, revenue and profitability of the company for the given dataset

[2] Berrish/Bullish Signals For Original Price

This is the important visualization. Here berrish signals basically tells that there has not been more sales over the short term thus reflecting the decrease in profitability over time yet suggesting a requirement for optimal pricing strategy.

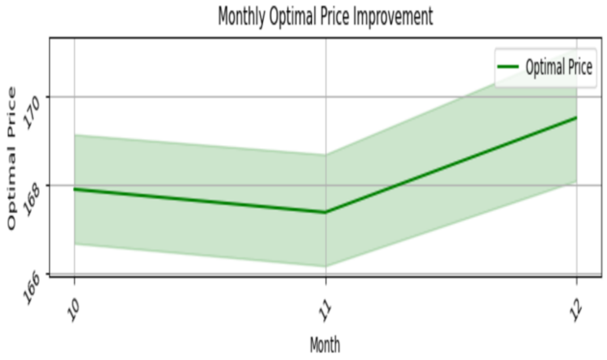


There can be several reasons for this trend. Some of them could be : Holidays / Promotions, A Mix Of Seasonal Shopping Patterns, Better quality products over offline shops or other ecommerce platforms, Shipping cost, Unoptimal prices of products, Poor quality of products, Unavailability of particular product in stock and many other possible reasons could be out there. This project's goal is to work address one of the problem, that is the *Unoptimal Price Of Products*

(B) <u>VISUALS AFTER DEVELOPING MODLES</u>

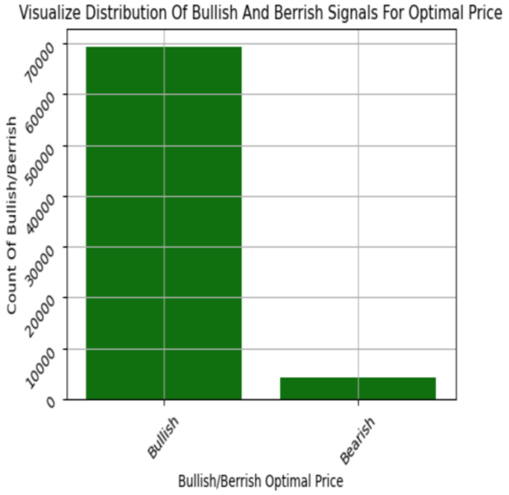
Upon implementing hybrid model and using it's results with other key features, I was able to obtain Optimal Prices. The visulas after implementing the hybrid model and obtaining Optimal Prices are below thus showing the improvement in sales and overall revenue as well as profitability

[1] Monthly Optimla Price Visualization



Thus, this visualization afer adapting Optimal Pricing strategies depicts the improvement in sales that will eventually impact the overall sales and revenue throughout the year thus achieveing the secondary goal of this project along with the primary goal

[2] Berrish/Bullish Signals For Optimal Price



Here bullish signals are greater than berrish signals, that basically tells that there has been large amount of sales over the short term thus reflecting the increase in profitability over time yet suggesting the improvement after implementing the Optimal Pricing strategies

Limitations : There were several limitations throughout carrying forward this project which were effectively resolved by using some tactics. Below were few of the limitations :

[1] <u>Absence Of Datetimstamp Feature</u> : The datetimestamp feature that basically gives the information of date which is crucial for developing TimeSeries model was not present initially within the provided dataset. This feature has been added by performing feature engineering before developing Time Series model.

Method Of Resolution :

Where the given information is, data has been scrapped from October 2023. Following feature engineering has been performed in order to obtain arbitrary datetimestamp for the Time Series model.

(1) <u>Year</u> : Random choice for year 2023 where the number of values is equal to the length of the dataset

(2) Month : Random integer between 10 to 13 where the number of values is equal to the length of the dataset

(3) Day :

(a) First created array of days in month whose datatype is dictionary. Thus, key October (10) to value 31, key November (11) to value 30 and key December (12) to value 31
(b) This array was mapped to Month feature giving the end result of days in particular month

(4) Date : Combined the Date, Month and Year features in order to get a single date

Note : The last date feature was optional and has not been used in any of the modeling but gives the clarity to the dataset in order to see which product was purchased on what date

[2] Memory Leakage : I encountered this issue when I was hypertuning the random forest regressor model. The primary reason was the output of ARIMA model had been passed as input to all the ML models. This hybrid hypertuned random forest regressor model was not able to handle the vast amount of data thus giving the **Pickling Error.** Pickling error occurs when there is memory leakage and the model can't handle the vast amount of the data leading to the crash of the model.

Method Of Resolution :

This issue was handled by the downsampling of the data. By downsampling the data, I mean taking only certain fraction of data so as to model can handle it and thus not crashing. I experimented with the certain fractions of data starting from 70% till the fraction the model can handle. After certain number of experimentations, downsampling the data to 40% of the data was the ultimate solution that model was able to handle.

[3] *Increased* Runtime : Upon implementing XGBoost Regressor model and hypertuning it, the runtime had increased. As we know, XGBoost Regressor model is an complex robust model as compared to other models. Since the output of ARIMA model was blended with XGBoost Regressor model as well as the hypertuned XGBoost Regressor model, the runtime of the model was increased.

Method Of Resolution :

In order to address this issue, I tried the light version of XGBoost Regressor model that is Light Gradient Boosting Machine regressor model which is popular for handling large datasets. This solution gave a good result. The runtime of the model was reduced for both the cases, that is with hypertuned as well as without hypertuned.

Thus, by opting these tactics and strategies I was able to resolve the problems and limitations I encountered to.

Results : Following are the results obtained upon implementing the hybrid model and using Optimal Pricing Strategies to obtain Optimal Prices

Aggregated Metrics Comparison (Price vs Optimal Price) :

Metric	Price	Optimal Price
Profit	$7.92 Billion	$42.99 Billion
Revenue	$9.32 Billion	$43.57 Billion
Inventory Turnover Rate	63.12	1806.84
Market Share	0.00	0.01
Popularity Score	3565.28	46982.67
Price Elasticity	2.33	6.55

This table summarizes the aggregated results of the improvements upon obtaining the optimal prices. We can see the substantial growth of profit from $7.92 Billion to $42.99 Billion and revenue from $9.32 Billion to $43.57 Billion. This result shows the overall increase in revenue and profitability as a result of obtaining Optimal Prices. We can even say these are more accurate since the results of best hybrid robust model was taken into consideration while obtaining the optimal prices. The increase in Inventory Turnover Rate from 63.12 to 1806.84 depicts the maximum sale of products throughout the year due to implementation of Optimal Prices which is ideally a strong positive indicator. The market share increase from 0.00 to 0.01 shows the recognition of the products in competitive market. Popularity score from 3565.28 to 46982.67 reflects the strong consumer satisfaction while the increase in Price Elasticity from 2.33 to 6.55 reflects the fact that any of the price change won't affect the sales of the product infact any price change would lead to increase in the profitability of the product. Said this, the implementation of Optimal Pricing strategy has imapacted the sales, revenue and profit of the company for the given dataset in a very positive which has been evaluated by the elevated values of other important metrics : Inventory Turnover Rate, Popularity Score, Market Share and Price Elasticity upon comparison of it's aggregated results with Price and Optimal Price

Conclusion / Discussion : Starting from the raw historic dataset with the result of degrade in sales throughout the year to the improvement of sales after implementing the optimal pricing strategies there has been strong improvement in overall profitability and revenue of the company which has been evaluated by improvement of the other key features as well, hereby meeting the project goals defined initially

Future Work : There is potential scope of future work : to meet the same project goals in a broader view. Below is step by step explaination for the same :

1. To integrate real time data from Amazaon website whose access is blocked by using API or building scrapper API for the same
2. In the raw dataset, there was limitation of DateTimeStamp
3. DateTimeStamp has been obtained by performing feature engineering that's mandatorily required for implementing TimeSeries model
4. All the models and the solutions are dependent on the arbitrarily obtained DateTimeStamp
5. If the DateTimeStamp comes with the real time dataset, adjust the feature engineering to remove the arbitrarily obtained DateTimeStamp
6. If the DateTimeStamp does not come with real time dataset, keep those and ammend the logic in the way to get random data for all the months, till current year and map the days to all the months according to number of days in a month
7. The current DateTimeStamp is the arbitrary DateTimeStamp feature engineered for months October, Novemeber, December and year 2023
8. Perform feature engineering to align the real time dataset features with the current dataset features in order to pass the same to existing designed models and optimal pricing strategies implemented
9. Since the hybrid model has been implemented, I faced the memory leakage issues. The current model and the optimal prices is dependent on the 40% of dataset used
10. To resolve the memory leakage issues (by implementing Deep Learning Models or other Robust Models maybe)
11. In the final updated dataset, clean the features that won't be required finally thus getting us a final clean dataset making it easy for the user to understand the goal of the project as well as the obtained results

Thus, by implementing these steps we can achieve same goals with the accurate results as upon the usage of best performing hybrid model in a broad view for the real time dataset that becomes unique with the strategies employed by the Amazon company in order to achieve the same goal as well as using the good coding standards of reducing line of codes

Source : Below is the source link from where I understood the strategies employed by the amazon in order to achieve the same goal

"Amazon Pricing Strategy Explained: The Ultimate Guide." *MetricsCart*, 18 June 2024, https://metricscart.com/insights/amazon-pricing-strategy/.