Expanding customer base for an Electronics Retail Store

End Term report for the BDM capstone Project

Submitted by

Name: Devansh Arora

Roll number: 23F2001431

IITM Online BS Degree Program,

Indian Institute of Technology, Madras, Chennai

Tamil Nadu, India, 600036

Contents

1 Executive Summary

Infinity Technologies, an authorized retailer of HP, provided a one-year set of 4,474 records for the four locations. Initially simple descriptive analysis gave some interesting findings. The sales obtained the highest level in April and showed an upsurge during the evening rush hour, especially between 6 PM to 8 PM. The correlation matrix was helpful in deducing the fact that the Product Level Discounts offered tend to decrease as the MRP increases.I was able to conclude that the Retail Notebooks group contributed maximum towards the sales, and also the products lying in the Consumer Volume Notebooks and Consumer Gaming Notebooks groups have a large scope of improvement in terms of sales. This was analyzed using the Pareto Chart.

Based on visit frequency, contribution to revenue, and similar statistics, the customer base was divided into three separate loyalty groups. Keeping top 6% customers as Platinum Group (Most Loyal) , the next 44% as Gold Group, and last 50% as Silver Group(New Customers) turned out to be the most feasible Loyalty Group Distribution. Each group comprised different purchasing behavior and, therefore, established grounds for tailored discounting schemes. These tailored schemes, developed for each loyalty group, were tested rigorously on a well-trained model for machine learning. The results were promising regarding an improvement in gross profit-a validation of this approach.

So, for any new customer who wasn't a part of the current dataset may be classified as a part of one of the 3 Loyalty Groups using a schema to give Loyalty Points to customers based on factors like Profit Generated, Longevity and Frequency of visit. Then, a specific set of discounts have been introduced for different loyalty groups which have been predicted to increase the gross profit generated by the store, using the trained models. Also, giving some high discounts on a few products like Laptop Bags and Stopwatch, when being sold with another high revenue earning product might prove to be beneficial in increasing sales as well as retaining customers.

2 Analysis Approach

Preprocessing and Descriptive Analysis

I received 1 Year sales data of their entire retail chain from Infinity Technologies General Store. This dataset was extremely expansive while also being very variable. The first task scrubbed the dataset away from the unnecessary features. In the variables fundamental to reasoning about a dataset, no principles were left out. In other words, the "RefPromoCode" pillar illustrations ought to be empty, illustrating that no promo code was applied when that product was being purchased. The cleaned data at the end consisted of 4474 rows and 36 lines.

The data was imported using the Pandas Library of Python. Some columns like ProductLevelDisc% and ProductLevelDiscAmount in the data had blank entries. These entries were filled with 0, because it was confirmed by the shop owner that any blank entries in these 2 columns meant that no discount was given. In other numerical features, any missing value was imputed with the mean of values present in that column. Also, some categorical features included CustomerType (NEW or EXISTING), ProductGroup etc. All these were handled using One-Hot encoding. This time, imputing was done based on the most frequently occurring value in the column.

The first detailed study was the basic reasoning dealing with calculating basic Mathematical boundaries such as Mean, Mode, Variance, Range etc. Using Google Colab Notebook graphs I generated graphs in order to better predict the dataset for generating meaningful observations. This was followed by a Time-sequence analysis to abstract inferences on consumer purchasing behavior during different months & time. Other than that, there were more histograms, bar graphs and strew plots used. These all were used to identify the sales trends across the different stores. A correlation matrix was generated between the columns Quantity, Profit, ProductLevelDisc, ProductLevelDic%, MRP and Amount. Correlation matrix was meant to give me an idea about the interdependence among these variables. The correlation matrix values range from -1 to 1, giving output in a normalized range. A positive value of correlation signifies that an increase in one variable, would lead to increase in the other variable as well. In case of negative correlation, the opposite is true, i.e. an increase in one variable leads to a decrease in another variable. Closer the numeric value of correlation coefficient is to 1, stronger is the correlation between the variables. Since the numbers are in a closed range, it becomes easier to compare them to generate meaningful insights. This makes it a better option for analysis as compared to other forms of similar analysis.

Loyalty Groups Division

To increase the Gross profit, the idea was to divide the customers into Loyalty Groups, and introduce specialized discounts. First, I figured out all features which could be used for grouping consumers right into clusters. With Python, I ran the K-Means Clustering algorithm to organize customers into loyalty clusters. Here, K-Means algorithm analyzed common features among different customers and hence, used these common features to group them into Loyalty Clusters. The customers were divided into a different number of clusters. It was understood that dividing customers into a large number of customers wouldn't be a feasible option because different discounting schemes cannot be run in parallel for a large number of different loyalty groups. Therefore, the number of clusters was set to 3. This number turned out to be ideal because out of the other values of number of clusters, lower value like 2 would be much less and doesn't provide a good enough distribution to work with, whereas for higher values like 4 or 5, the customers are distributed such that there are very less number of customers in one cluster, while a very vast majority is categorized in another cluster. These loyalty groups would be used to generate personalized discounts for different customer groups, which would hence attract customers and motivate them to stay loyal to the store.

Discounting Analysis

After this, the aim was to analyze the discounting schemes which would increase the Gross Profit. To do this analysis, I tried using several regression models like Linear Regression, Polynomial Regression and Random Forest Regression. A regression model helps in simulating/ making predictions based on the previously observed patterns from the input data. These regression models would be useful for predicting profits generated by the store given the discounting schemes that I introduce. I utilized the sk-learn library from Python to implement these models. The model was given all of the data (except Gross Profit and Quantity Sold) as input and the target variables were Gross Profit and Quantity Sold.

Also, to capture the importance of time of sale, the BillTime and BillDate columns were divided into 3 separate columns, namely Hour, DayOfTheWeek, IsWeekend. Hour captures the Hour of the day when product was sold, while DayOfTheWeek captures a numerical value highlighting what day it was when the product was sold. IsWeekend is a boolean type column wherein I have only 2 values, 0 and 1, representing False and True respectively.

Now, all the models were trained, with different Hyperparameters. Hyperparameters are the aspects of the ML model which are under control of the creator of the model, which can be altered to generate better outputs. The model which performed the best in validation was selected to be the final model. This was achieved using the GridSearchCV function in the sk-learn library. This way 4 models were trained. One model captures the features from the overall dataset, and the other 3 models capture only the features from the part of the dataset corresponding to one of the 3 loyalty clusters. Then, these models were run on test dataset, and the Root Mean Squared Error (RMSE) and R^2 Metrics were utilized to judge the models performance on test dataset. The Root Mean Squared Error metric is a measure of how much the model is diverging away from the actual values. If the RMSE is small, then we know the model is performing well. The R^2metric on the other hand judges how well the model is able to explain the variance in the dataset or how good the model is at extracting out the patterns from the data. This metric's values range from 0 to 1, meaning that the model is just not able to do the required task, and 1 meaning the model is doing the required task exceptionally well. However, the metric's value can also be negative, stating that the model is giving extremely poor performance, or better stated as, the model would have done a better job than this had it been predicting the average each time the model was run. These metrics judge the disparity between the true values and the values model predicted. A 'Residual Graph' was plotted, indicating the difference between the actual value and predicted value of Gross profits in the test dataset. The X-axis is the actual Gross profit generated, and the Y-axis shows the difference. These graphs would show the difference between Actual Profit generated by the store and the profit values predicted by our model, when the model was being tested on data whose results were already known. This would be helpful in judging a model's performance, as we would know in what range of profits the model is giving a good output by looking at the deviation from Y = 0 line (shown as red dashed line in the plots). The Y = 0 line shows the ideal case, where the model predicts exactly as the real value of profits obtained. Also, a bar graph of different feature importances for each model was plotted to check what features contribute the most to the model outputs. This would be indicative of what factors about the product affect the customer's choices the most.

Once the models were ready, an algorithm was run using those models. I first introduced a list of discount percentages, which included 5, 10, 15, 20, 25, 30, 35 and 100. Now, in one cycle one of the products was given a discount value while all other features remained the same for all products, and the models were run to predict the new gross profit generated and hence calculate the percentage increase in the gross profit. This procedure was repeated for each product and each discount value, and the products with top 10 values of percentage increase in Gross profit for each of the discounting values. This was meant to give me an idea about the products which lead to the maximum profit generation for each of the given discount percentages within different loyalty groups along with overall dataset analysis. These findings were then used to introduce the final discounts for the stores.

3 Results and Findings

Initial Statistical Findings

The following correlation heatmap shows a slight negative correlation of Product Level Discount Percentage with MRP, Amount and Profit. A negative value in correlation heatmap portrays that as the Amount and Profit increase, the Product Level Discount Percentage offered seems to be decreasing. This means, at a high MRP the customers purchase the product even at a low discount. However, this correlation is very weak, as deduced from the small numeric values written in the heatmap. This means, profit directly doesn't have a very high effect on the gross profit. However, the Product level discount percentage is directly linked to Amount, as the Amount is calculated after applying Discounts and Taxes on the MRP. Though MRP and Amount themselves have a low correlation with discount, they have a significantly high positive correlation with the Profit (0.90 and 0.94, respectively). This implies that the Profit generated is highly dependent on MRP and Amount, and hence slightest of changes brought to these variables, would generate a significant change in profit. And this slight change can be brought in by introducing new discounts.

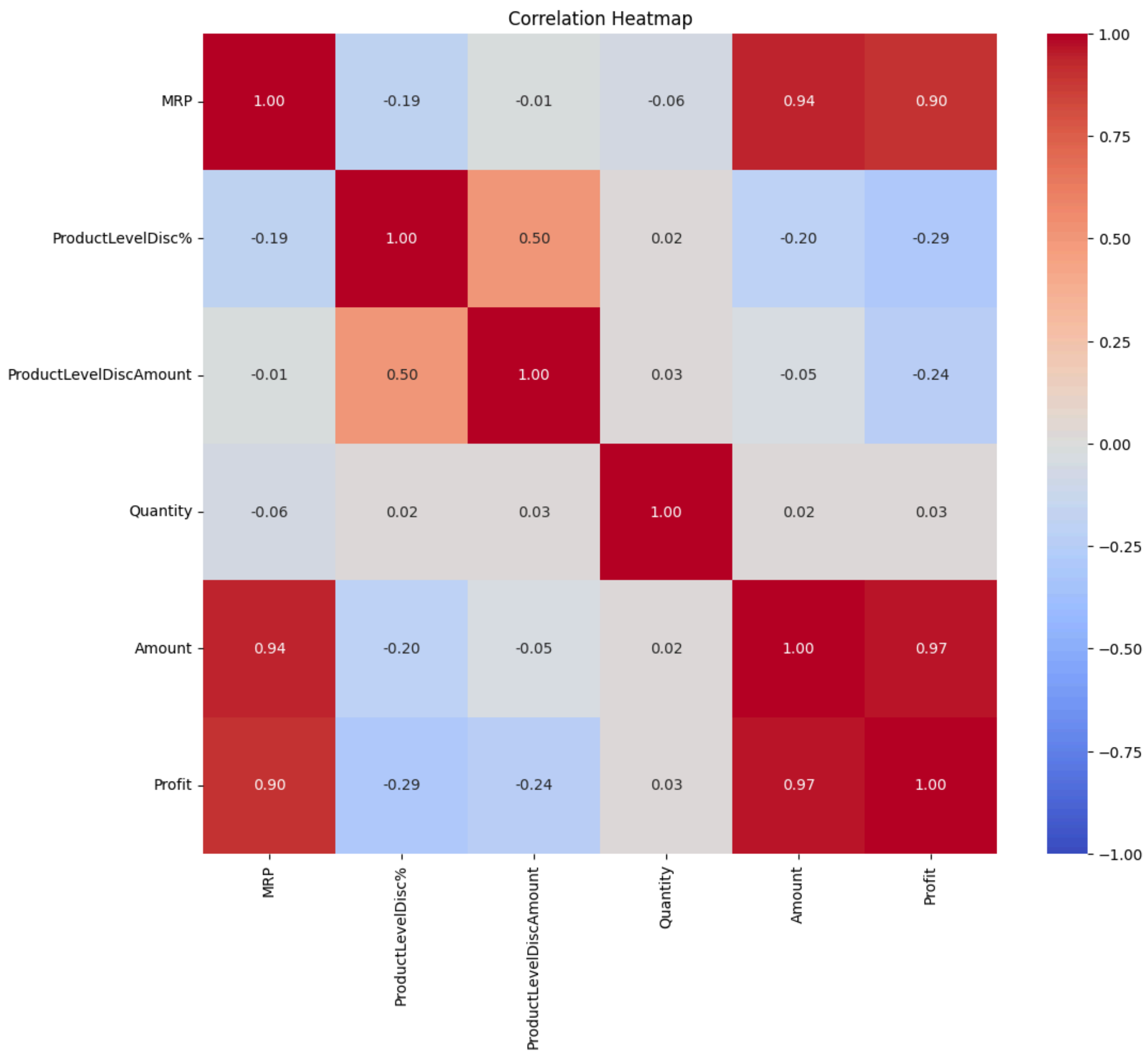


Figure 1 : Correlation heatmap between different numerical variables.

From the data, we can note that Mean MRP of products sold in the retail store comes out to be 31043.88 rupees. I detected the outliers, using the InterQuartile Range using the following data:

Quartile 1 : 10169.49 (Lowest value among the bottom 25% values of MRP)

Quartile 3 : 50000.0 (Highest value among the top 25% values of MRP)

This gives the percentage of outliers to be 1.30%. Any outlier is not considered while doing the initial analysis. However, incase of Loyalty Groups and DIscounting Analysis, no datapoint can be ignored since each customer and each product holds a value. Therefore, the whole dataset is taken into account at those stages of the analysis.

Next I have plotted a line chart for the number of customers who bought from the store at different hours of the day in Figure 2 . This is meant to be used to analyze the variation in rush at the stores as the day passes.

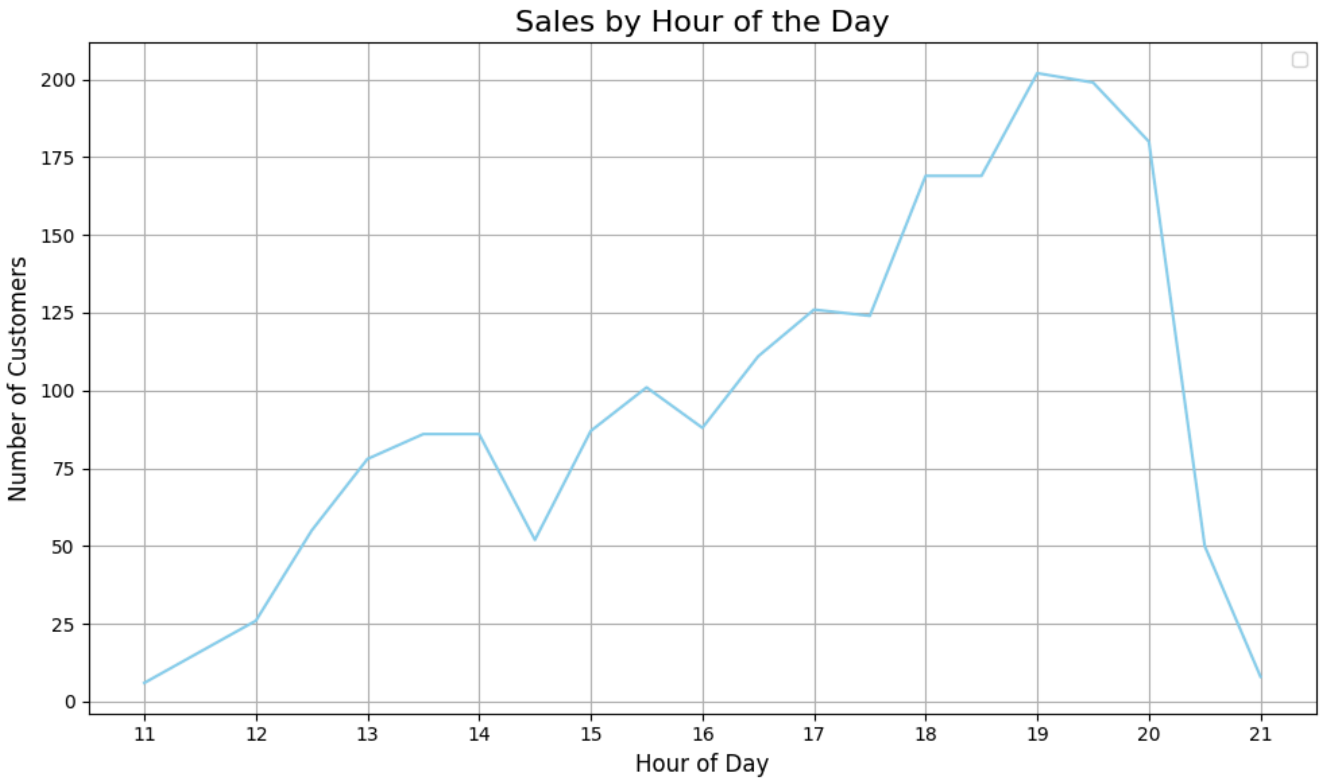


Figure 2 : Line graph showing the sales at different times of the day, taken as a cumulative data from the 1 year sales dataset.

We can see that evening 4:00 pm to 8:00 pm is the high time for sale of products in the store. This tells us that we need a higher workforce around these hours of the day. For the rest of the day, a part of the workforce may be utilized for tasks other than those of a salesman. Also, we may infer that the time of the day may form a contributing factor towards sales of products. This information is useful when identifying the features to be selected for generating a model for discounting analysis.

The following histogram in Figure 3 shows the distribution of sales (quantity of product sold) with MRP. It can be observed that maximum sales occur on lower MRP products. We can see that maximum sales occur in the MRP range of 0 to 3600 rupees. Then we can see another rise in the 10785 to 14380 rupees bin. This retail store forms a prime spot for laptops and printers. As a result, sales are high around laptop and printer prices. Nonetheless, a lot of inexpensive items, such as smartwatches, wireless mice, and HP laptop bags, are given away for free with other items. Because these products are sold with other products, this also results in an increase in the bars that represent their MRP.

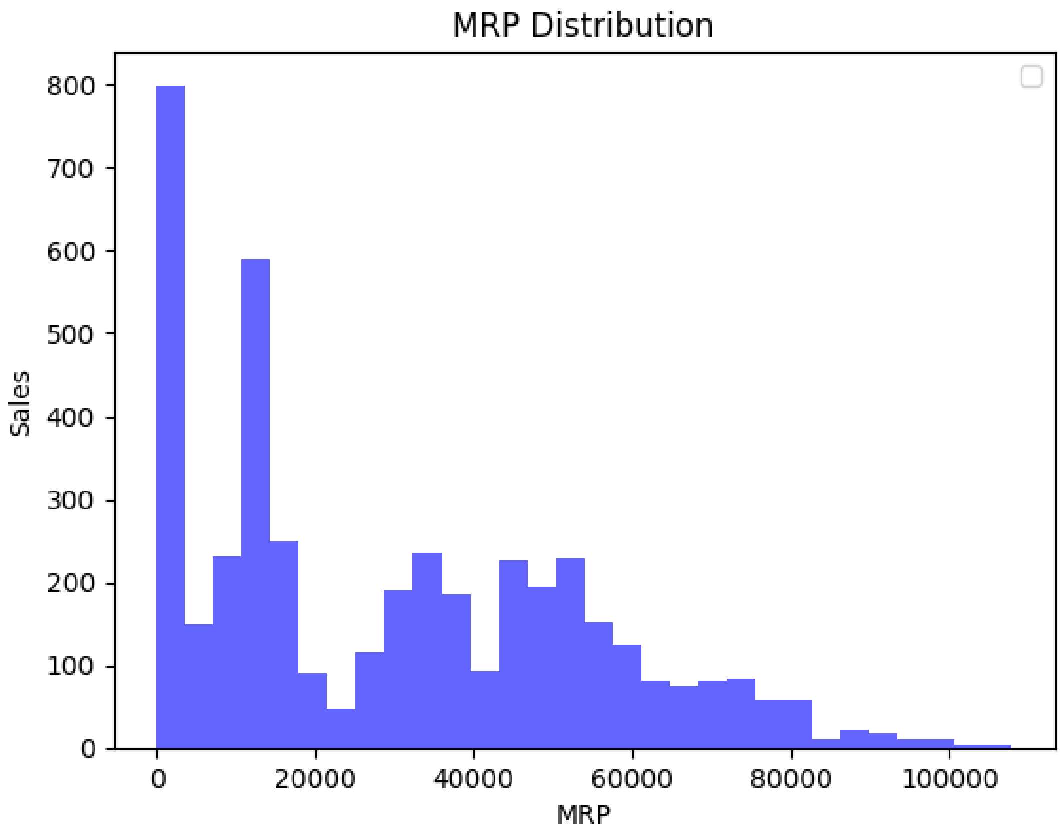


Figure 3 : Histogram between the Sales (or the Quantity of products sold) vs their MRP (in rupees)

The correlation coefficient between MRP and Sales is -0.67 showing a moderate to strong correlation between the two variables. This tells the price sensitivity, wherein higher priced products are sold in lesser quantities as compared to low priced products. Therefore, we can say MRP is a contributing factor in judging the sales, but not strong enough to explain the trends well.Also, we must note that among the 375 products being sold at the store, the most sold product turns out to be Printer LaserJet Professional P1108 - HP.

We can see a Pareto Chart in Figure 4 between the Product Groups and Sales Amount. Here sales amount refers to total Amount/Revenue generated through sales of products in each of the given Product Groups. The blue vertical Y-axis shows the Revenue on a scale of 1e7 i.e. a unit increase in the vertical Y-axis corresponds to change of 1e7 in Revenue. The blue bars correspond to the value of Sales Amount for the Product Group. On the right side, we have another orange Y-axis, corresponding to the Cumulative Percentage contribution of the product group in the total sales. The cumulative percentage implies that it represents the percentage contribution of the current product group under assessment along with the previously assessed product groups. For example, we can see Retail notebooks correspond to a value of 50 on Cumulative Percentage Scale, which means retail notebooks alone contribute to 50% total revenue generated. However, the corresponding value for Consumer Volume Desktop is around 65. This implies that Consumer Volume Desktops, in addition to Retail Notebooks, contribute to 65% of total revenue generated, which means that Consumer Volume Desktop alone contributes to 15% (65 -50 = 15) of the total Sales Amount.

The product group generating maximum sales and profit turns out to be Retail Notebooks. However, the most selling product isn't a part of the group generating maximum sales. This means the Retail Notebooks group products generate decent sales which overall have a significant contribution in the gross profit. We can use this knowledge to our advantage when assessing discounting schemes which may turn out to be beneficial for the store.

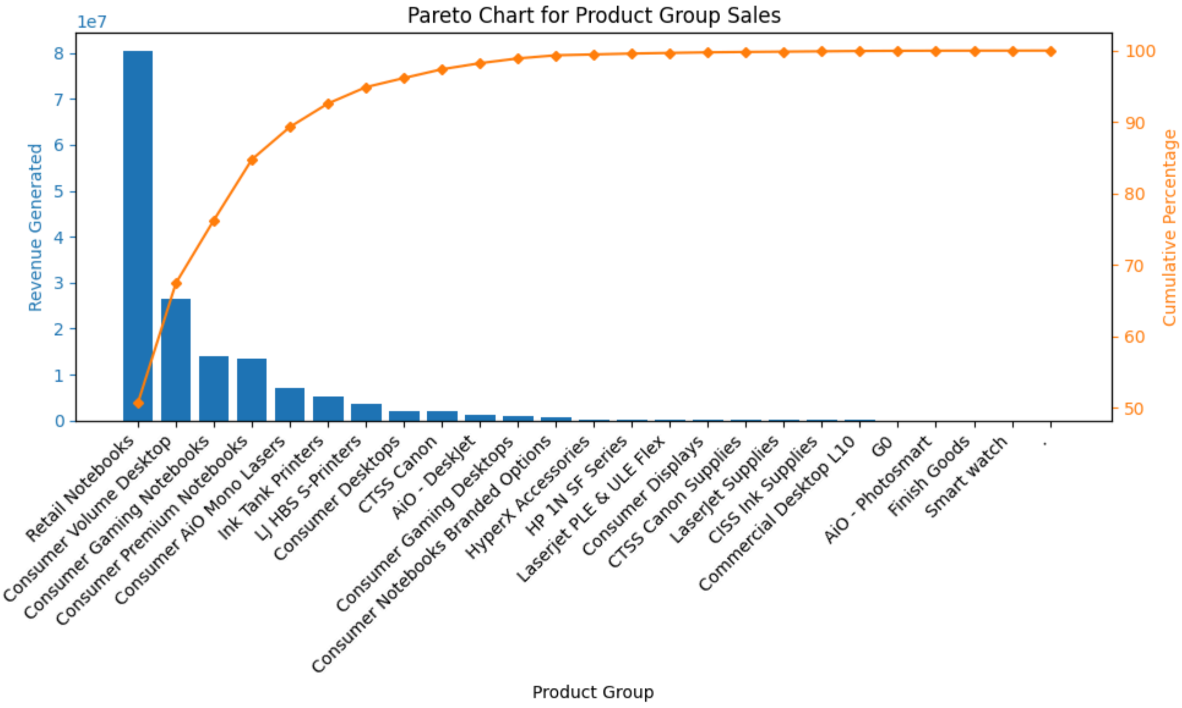


Figure 4 : A Pareto Chart (Orange) of different Product Groups present at the store with respect to percentage of contribution to sales, along with a bar graph showing Revenue generated (on scale of 1e7 rupees) over the whole calendar year.

We can also see from the Pareto Chart that Consumer Volume Desktop and Consumer Volume Notebooks are respectively the second and third most selling category. However, The sales of Consumer Volume Desktop is less than half the sales volume of Retail Notebooks. This tells that improving the sales of Consumer Volume Desktop and Consumer Gaming Notebooks should be worked upon. These categories show a high scope of improvement in terms of sales.

Loyalty Group Division

Loyalty group allotment file : [loyalty_clusters.csv]

The customer base is divided into 3 Loyalty Groups. Analysis of different loyalty clusters gives the following insights:

I have used box plots (Figure 5) to visualize the range of values covered under the span of each cluster. The top and bottom edge of the box correspond to the 1st and 3rd quartile. The line in the middle of the box represents the median of the dataset corresponding to each of the clusters. The additional extensions above and below the box represent the limits of the Inter-Quartile Range. Interquartile Range is a mathematical range of values used to pick out outliers from the dataset. Any value lying outside this range is considered as an outlier and is represented as a hollow bubble, as in case of the box plot of cluster 2.

The 0th cluster shows a very small box plot. The average number of visits for the customers in cluster 0 comes out to be 1. An average of 1 is only possible when all customers in cluster 0 have visited the store only once, since the number of visits can only hold positive integral values.

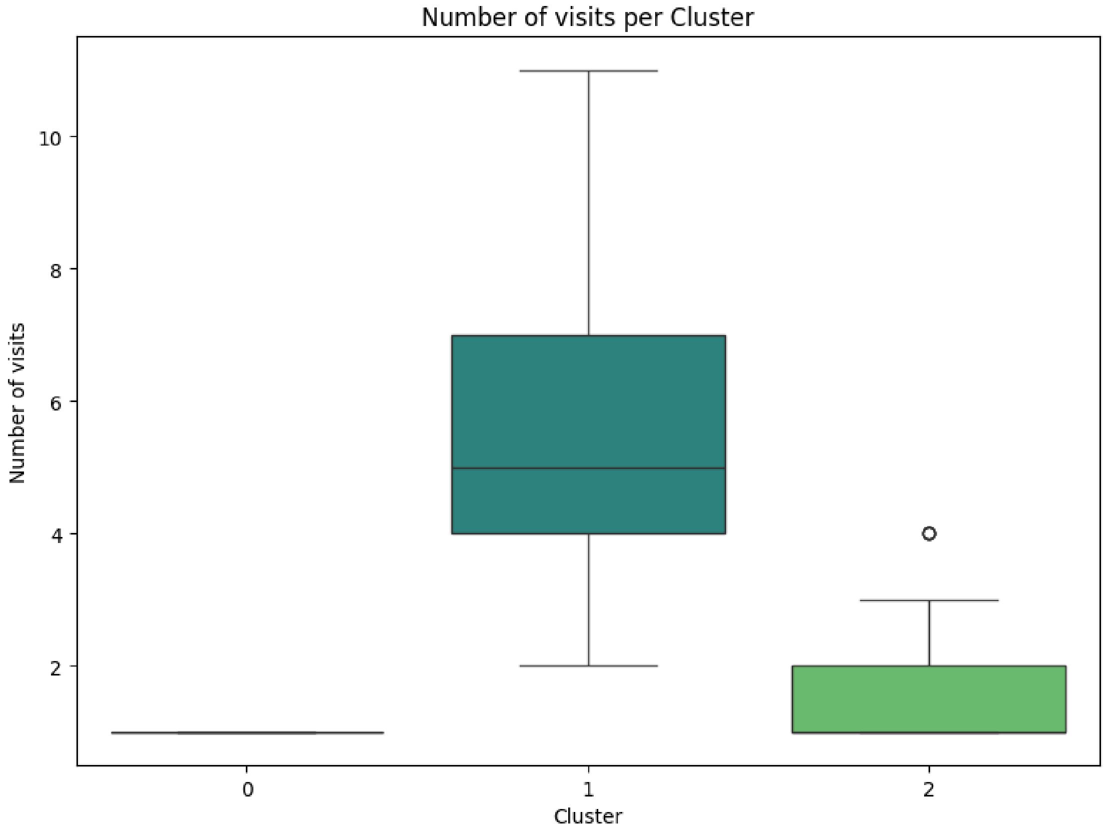


Figure 5 : Box-plots of Number of Visits of the customers in a given cluster for each of the 3 clusters.

Cluster 1 shows a large boxplot. The average number of visits for cluster 1 comes out to be 5.377. This means, this cluster consists of very frequent customers. Cluster 2 has an average of 1.35 visits per customer. These customers are, then, not very frequent visitors but they still maintain contact with the retail store.

In cluster 0, as I already mentioned, each customer visited exactly once, and hence their Longevity is 0. Cluster 1 shows average longevity of 245.16 days. Cluster 2 shows average longevity of 20.44 days. This means that the people in cluster 1 were visiting more times and for a longer period of time.

Also, cluster 1 contains 61 customers, cluster 2 has 449 customers and cluster 0 has 497 customers.

The violin plot (Figure 6) is a density distribution of data points in different clusters across ranges of profit. Wider the violin plot, higher is the density of data points lying in that range of Profit. We can see a highly widespread violin plot for cluster 1 while those for clusters 0 and 2 are very shrinked.

The profit distribution in this violin plot clearly depicts that maximum profit is generated from customers in cluster 1. The average profit generated by cluster 1 is 163094.35 rupees. That generated by clusters 0 and 2 are 9067.78 rupees and 19590.16 rupees, respectively.

Therefore, we can say that there is a small group of people in cluster 1, which contribute to a large segment of the gross profit and also stay connected with the store. They form the set of the most

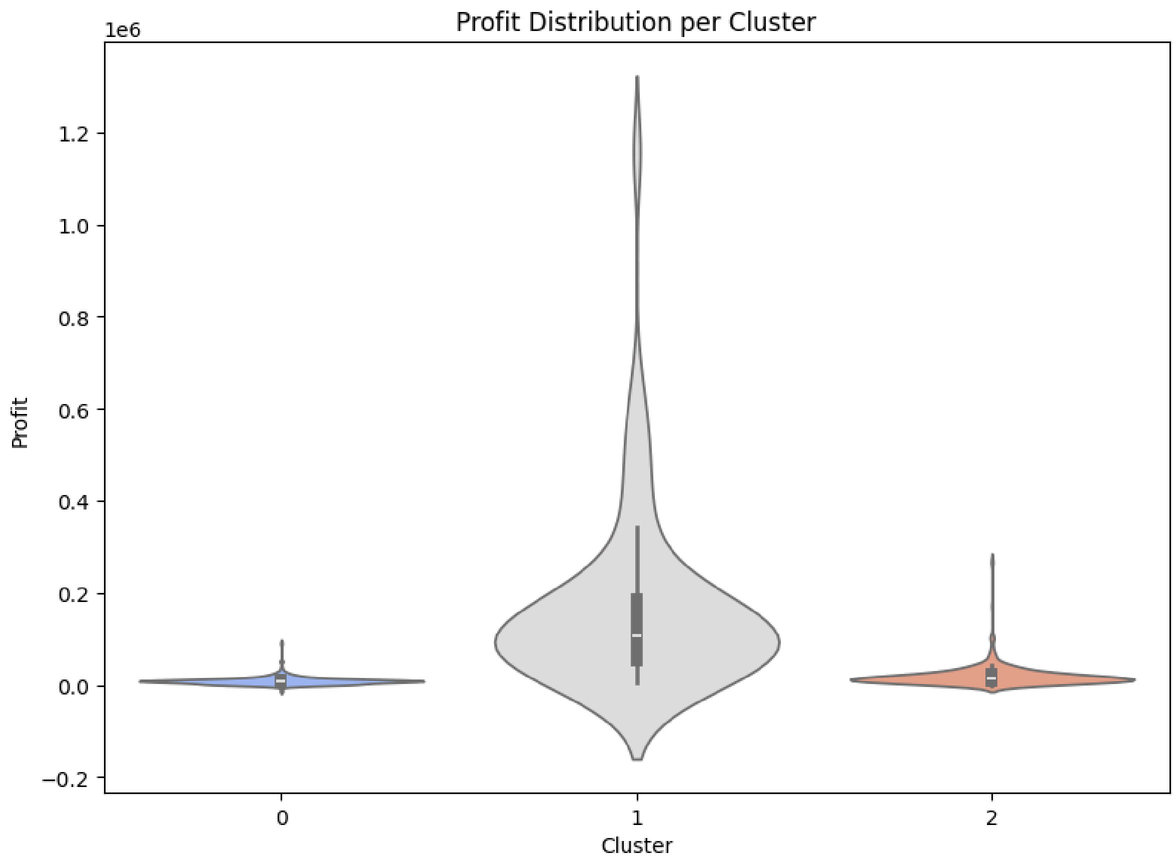


Figure 6 : Violin plots of Gross Profits generated (on scale of 1e6 in Rupees) from the customers in a given cluster for each of the 3 clusters.

loyal customers of the store. The customers in cluster 2 stand second on the loyalty pedestal. They are not as frequent visitors but they still generate significant profit.

It can also be seen from the file that cluster 0 consists of all NEW customers. The EXISTING customers are distributed between cluster 1 and cluster 2. Somehow, we know that if a new customer came into the store during the year and that same person returned again to the store, that person was tagged as Existing Customer. Cluster 1 and 2 both have only existing customers. This means, cluster 1 and 2 include both kinds of customers, those who were already existing customers before that year and those who came as new customers but decided to return to the same store again within the given year.

Hence, we can divide loyalty groups as follows:

Platinum : Cluster 1 : The most loyal group

Gold : Cluster 2 : The second most loyal group

Silver : Cluster 0 : The least loyal group

Discount Analysis Model

Test output for the best models : output_file.log

This log file contains the best parameters of the models trained to predict the gross profit obtained. For each model, it gives the value of Root Mean Squared Error (RMSE), which shows how much error occurs between the gross profit values predicted by the model and the true gross profit values. Lower the error value, we know better is the model performance. We also have R^2 metric, which is another way of knowing how well the ML model performs. A metric value greater than 0.95 tells that the model is performing very well, that is, it is highly effective in simulating the profits obtained by store given the required parameters like the customer data, the information about the products being sold etc.

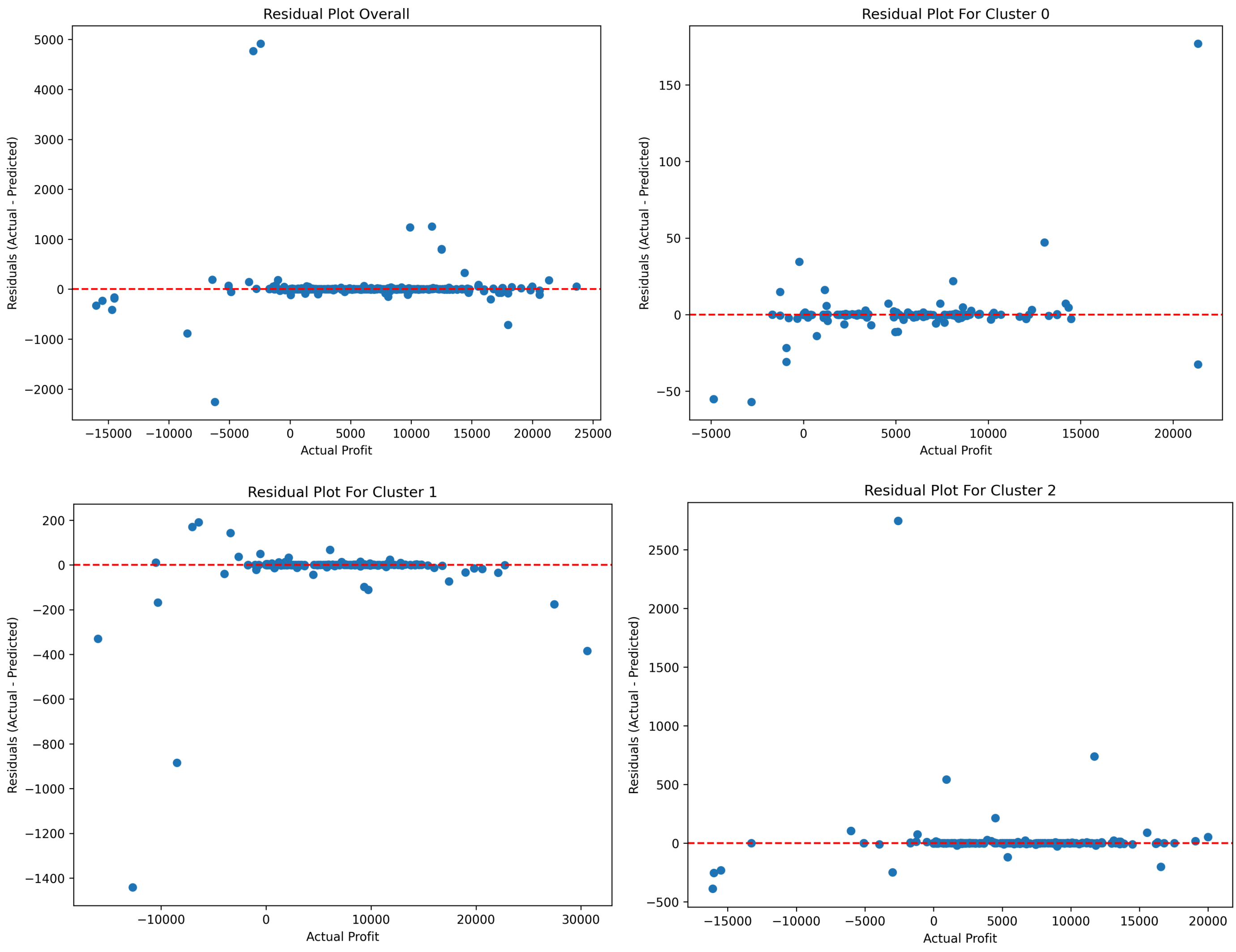


Figure 7 : Residual plots for predictions made by the 4 trained models : Overall Model (Top Left), Cluster 0 Model (Top Right), Cluster 1 Model (Bottom Left) and Cluster 2 Model (Bottom Right)

This confirms that the trained model can effectively predict the gross profit generated by the store, and we can use it to judge the effect of introducing new discounts on the products. The model would be able to give us an idea that given a discounting scheme, would it result in an increase in gross profit over a period of one year or not. 4 models are made for 4 different groups. The first model named 'Overall" is made using the complete dataset. The other 3 models are made using dataset filtered using Loyalty Groups. First, the original dataset was divided into 3 parts. Each part corresponded to customers in different loyalty groups, that is, sales data of all customers in one loyalty group was separated from the rest of the data. Now, using these 3 new datasets, models were trained, and outputs were recorded in the output_file.

I have used residual plots (Figure 7) to the variation in the predicted output value from the actual expected value. The Residual Y-axis shows the difference between the actual and predicted values of profit. The X-axis varies in the actual value of the profit generated. We can observe that in all the residual plots, the maximum density is around the red-dashed line, which is the Y = 0 line. This means that values predicted by the models are close to the actual values mentioned in the dataset. Therefore, we can say that the models work reasonably well. This can also be verified from the metric values obtained from testing the model.

The residual plots show that there is not much difference between the actual and predicted values. However, Cluster 1 has a narrower spread of residuals around the red line (that is, zero value) than Cluster 2. This shows that Cluster 1 has relatively better model fit than Cluster 2. Cluster 2 shows higher variance in residuals. Overall, the models are fairly robust. Also, we can note some negative profits being projected in the residual graphs. These negative profits correspond to the deals wherein the customers returned some product.

We can see a varying range of values over the X and Y axes of the 4 plots. This shows that the 4 models are operating in different ranges of profits. In case of cluster 0 , the least value of profit goes to -5000, while for cluster 1 it goes even below -10,000. As the profit becomes more negative, or we can say the loss increases, largely negative values of residual can be seen in case of cluster 1. This means that the actual value of profit is much lesser than the predicted value of profit. However, if we observe the axes carefully, we can note that the residual is just around 10% of actual profit incase of cluster 1 for points below the red line. This means that still the predicted value is somewhat close to the actual profit value, and still the predicted value is negative. This means that in these cases the model was correctly able to identify its case of a loss, though it missed from the actual value by an error of 10%. Thus, we can conclude the model is robust in classifying if the given deal would lead to a profit or a loss. In other cases as well, we can see similar trends, except the fact that the residual is not around 10% of actual value of profit. However, if the model made an error in judging if the deal would lead to a profit or loss, the residual value would be less than the actual value of profit incase of negative profit (or more than the actual value of profit, incase of positive profit). And this is almost never the case, as visible from the graph.

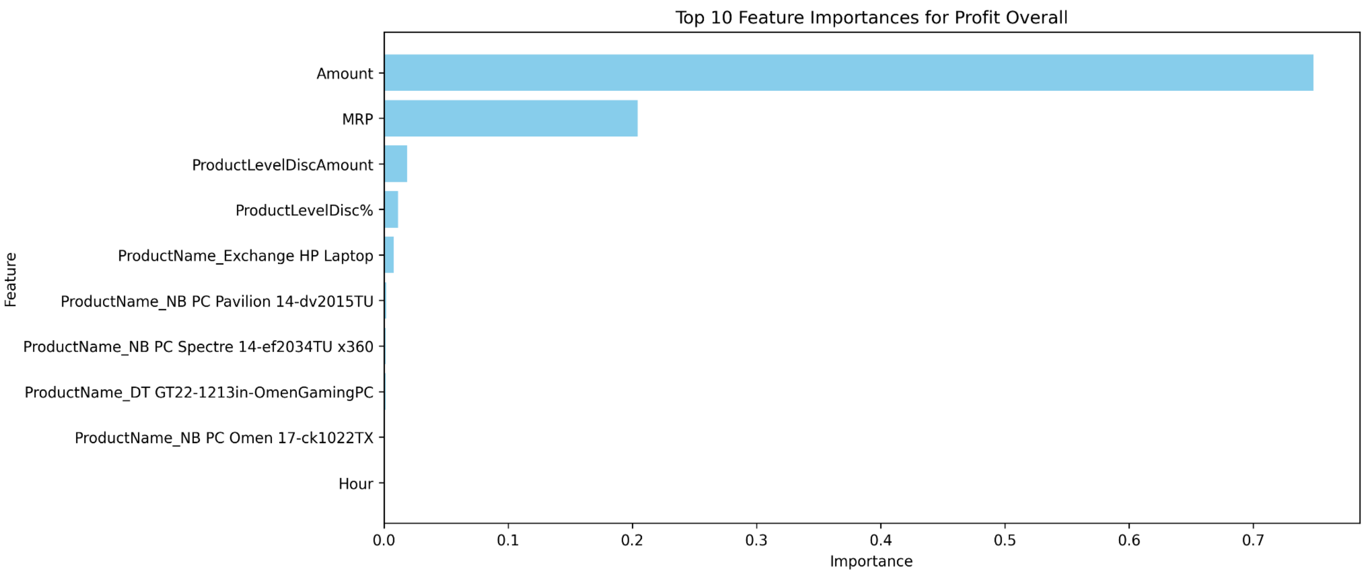
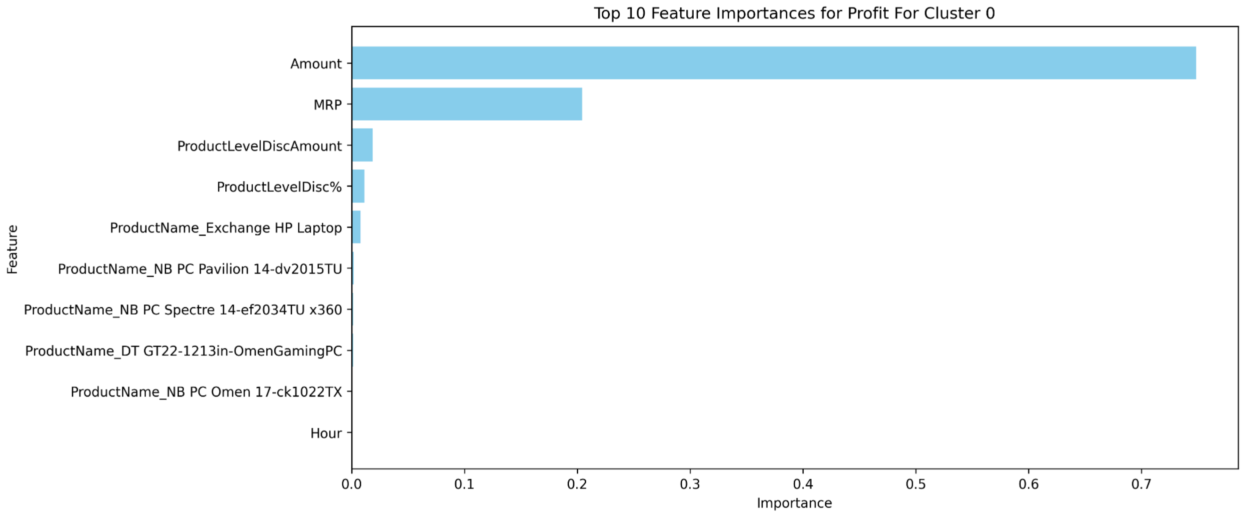


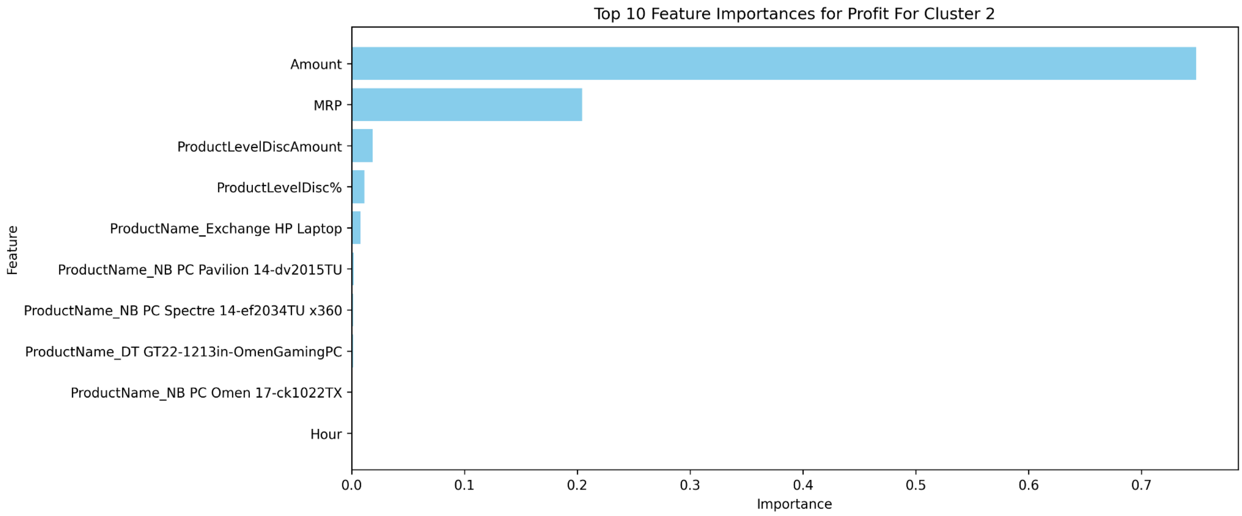
Figure 8 : Bar chart of the top 10 most important features used to judge the profit generated, as extracted from the trained Overall model

The bar graphs in Figure 8 and 9 show the importance of 10 most important features used for predicting the profit generated using the overall model. The vertical Y-axis corresponds to the Feature and horizontal X-axis corresponds to the fraction of importance of the feature under consideration. Fraction of importance means, out of the total importance of all features, what 'value' of importance the given feature holds.

In Figure 8, we can observe that the final amount at which a product is sold , named as "Amount" in the bar chart, forms the most important feature. We may also note that the importance of the most important feature(Amount) is almost 3 to 4 times that of the second most important feature(MRP). This shows that the most important feature has significantly very high importance. Thus, a small change in the Amount would show a significant change in sales. However, it can be seen that the value of product level discount percentage doesn't show a lot of importance. It shouldn't be forgotten somehow that Amount and Product Level Discount Percentage are not independent, and hence any change in Product Level Discount Percentage would still bring a significant change. Similar trends can be viewed in the case of other 3 models as well (Figure 9).

We may notice that the ProductName_Exchange HP Laptop is seen to be having somewhat significance in all graphs. It can be because Exchange HP Laptops correspond to the Laptops which were returned to store. These deals always lead to a significantly large value of loss, and this name becomes an important factor for the model to predict if the deal would end up in a profit or a loss.





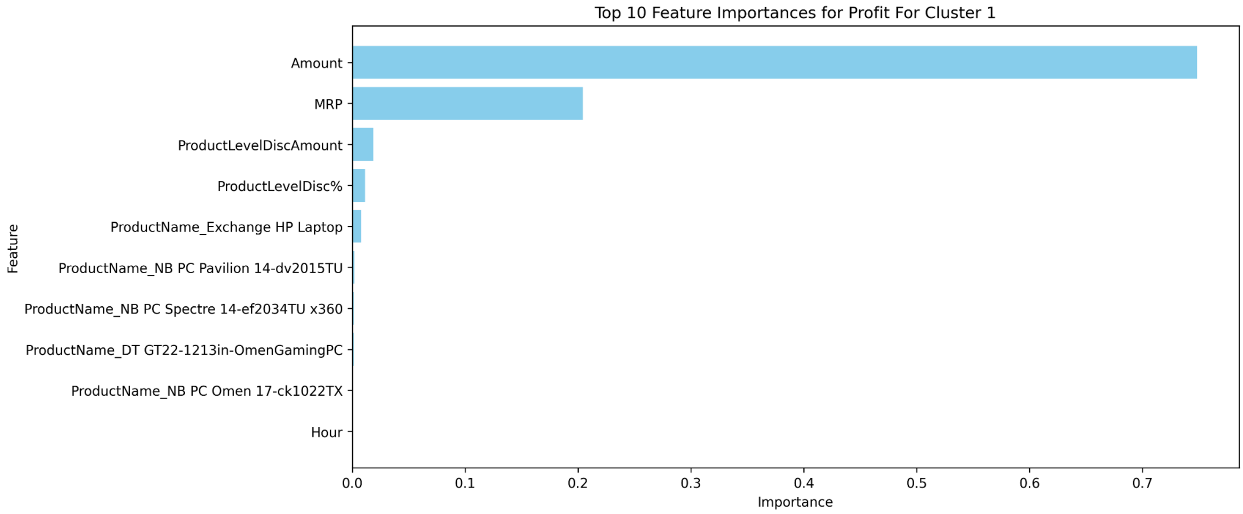


Figure 9 : Bar charts of the top 10 most important features used to judge the profit generated, as extracted from the trained Cluster 0 model, Cluster 1 model and Cluster 2 model

Model Outputs Analysis

I used these models to simulate the functioning of the shop, in terms of the profit generated. The discount_file shows that at a given discount percentage, what products would generate the maximum discount. They are arranged in order of percentage increase in the total gross profit.

Outputs on different discounting schemes: discount_file.log

It could be inferred that, irrespective of the loyalty group, the first thing that influences the buyer is the MRP, and then once the discount is introduced, that final amount gains higher weightage. However, we do not observe much dependence on the Discount Percentage or Discounted Amount values. This can imply that though it captures attention, Discounts are not specifically something that the customer is always looking for. Their initial judgment is done via the MRP, and later adding the discounts strengthens their will to buy the product. It can be seen in the discounts output file, that individually introducing discounts on products don't lead to a significant increase in the profit percentage. However, it can also be seen that introducing 100% discount for some products leads to a growth in profit, which can be implied from the fact that when buying an electronics item the customers are looking forward to getting some other niche products like a Bag or a Smartwatch for a lower rate, or free.

But when it comes to selling the products like Bags and Smartwatches standalone, they may not be discounted. It can be seen from the output file that the percentage increase in profit with or without discounting is nearly the same till 5% discount. This can be used as a selling strategy in such a way that these products may not be discounted initially, but the seller may add a discount of 5% if they feel the need to.

We can see from the output file how different products lead to an increase in profit for different products. The output file lists the top 10 products based on the % increase in total Gross profit generated by the store, given that they apply the chosen profit on the chosen product.

Somehow, we can note that at many places, Exchange laptops come up in the top 10 items which lead to significant increase in profit. However, it should be noted that imposing a discount on

Exchange laptops isn't a feasible option, as incase of Exchange Laptop, the customer comes to return an already bought product.

We saw that Printer LaserJet Professional P1108 - HP was the most sold product, but however in places where its name occurs in discounts_file.log, it shows a negative growth in profit. This gives a learning that a product giving good sales even without a discount shouldn't be fiddled around with. It might have the opposite impact of what one may expect. We also noted that Retail Notebooks form a significant contributor towards the gross profit. We can see some products belonging to that group in the discounts_file.log, generating a positive value of percentage increase in profit. This implies that these moderately well selling products may boost up the sales when supported by a good enough discounting scheme.

We may also note that products whose name starts with DT AIO occur very often in the file. These products belong to the Consumer Volume Desktop group, which was 2nd most contributing group as per the Pareto Chart. There as well, we were able to deduce that promoting products of this group by giving discounts could prove to be a beneficial idea. This is validated by the observation we made about frequently occurring products.

4 Interpretations and Recommendations

Generic Recommendations

a. As observed from the line graph on Sales by Hour of the Day, the shop may manage the workforce according to the requirement at a particular time. Keeping the whole workforce at the store for the full working hours of the shop might not be as useful. The major workforce needs to be active from 4:00 pm to 8:30 pm as they are the peak sales timings. Rest of the time, the work force may be judiciously used for other tasks like inventory management.
b. Higher focus may be given on the sales of products belonging to the Consumer Volume Desktop, Consumer Gaming Notebooks and Consumer Premium Notebooks groups. They show a chance of pushing off the sales, as discussed with regards to the Pareto Chart for Product Group Sales.
c. Products like Laptop Bags or Smartwatches may be given as a free gift along with some other high profit generating deal. The discounting and giveaways need to be balanced and managed as per the product demand.

Loyalty Groups

The first step is to divide customers into Loyalty Groups. The scheme which may be used to do this is as follows:

a. Frequency Based Points:

   Customers are first given points based on their frequency of visit to the store.

   High Frequency (top 35%): 10 points

   Medium Frequency (35%-65%): 6 points

   Low Frequency (bottom 35%): 3 points

b. Profit based Points :

For every 150 rupees profit they contribute to, the customer is awarded 1 point.

c. Longevity Points:

Longevity is defined as the number of days between the first visit and last visit of the customer.

For every 75 days of a customer's relationship with the company, 1 point is awarded.

Total Points = Frequency Points + Profit Points + Longevity points

Now using the points assigned to the customers, they can be divided into 3 loyalty groups:

1. Platinum : Top 6% customers : Represents the smallest but the most elite group of customers.
2. Gold : Next 44% customers : A group of customers with significant contributions to the store's profit.
3. Silver : Rest 50% customers : The largest group with customers having the lowest contribution to the store's profit.

This schema is designed to match the proportions of customers in original Loyalty Groups (in loyalty_clusters.csv), while maintaining logical and meaningful divisions.

Therefore, for any new customer who wasn't mentioned in the given dataset till yet, they may be assigned the Silver Loyalty Group, and then from thereon their Loyalty Group may be updated based on their future visits to the store.

Discount Schemes

Next, some special discounts can be introduced for different loyalty clusters.

PLATINUM GROUP :

- NB PC Pavilion 14-ec1003AU : 5%
- NB PC HP15s-eq2144AU : 5%
- NB PC Pavilion 14-dv2015TU : 10%
- DT AIO 27-ca1142in-Black-PavilionTS : 10%
- DT AIO 22-dd0480in-SnowWhite-Pavilion : 10%

GOLD GROUP :

- NB PC Pavilion 14-dv2015TU : 5%
- NB PC HP15s-fq2717TU : 5%
- DT AIO 24-cb1237in-StarryWhite-Pavilion : 5%
- NB PC ENVY 13-bf0121TU x360 : 5%

SILVER GROUP :

- NB PC Pavilion 14-dv2015TU : 5%
- NB PC Pavilion 14-ec1003AU : 5%

These discounts have been introduced keeping in mind that they lead to net increase in the profit in the long run. This way we aren't compromising on the profits earned by introducing new discounts. These specialised discounts would attract customers and motivate them to stay loyal to the store. This way it would serve our prime purpose of expanding the customer base of the store.

Along with such discounts, giving away a few products for free also looks like a workable scheme as has been observed in the dataset. These giveaway deals might as well be altered as per the loyalty groups. However, it may be noted that giving more giveaways to Platinum Group has better outcomes as observed from the analysis. For other groups, along with the high revenue generating products, these products may be given at a highly discounted rate.

We may also note, that when someone comes to buy products like Laptop Bags or Smart Watches, they may be given without any discounts or maybe 5% discount if the customer demands some, or if the seller feels the need to take this step to increase the sales.