High Profit Prediction

MLDS 400 Group 13
Lisa Gao, Inu Tenneti, Gace Xie, Oliver Zhou

Executive Summary

We aim to analyze whether the store achieves high or low profit across various brands, stores, states, and years. We utilized binomial logistic regression, KNN, and Decision Tree models to predict 'High Profit' and 'Low Profit', considering factors such as state, store name, brand name, selling price, quantity, original price, cost, retail price, and discount rate.

Business Case Study

In this project, we explore the Dillard dataset to decipher the nuanced factors driving retail store profitability. Beyond conventional metrics like store names and locations, our focus extends to critical financial variables such as sell price, quantity, original price, cost, retail price, and discount rate. By delving into these intricacies, we aim to uncover hidden insights into how these metrics collectively influence the profitability dynamics of retail stores.

Our analysis spans individual stores, diverse brands, states, and time frames, providing a comprehensive view of the factors shaping retail profitability. The outcomes of this study are poised to offer actionable insights for retail managers, brand owners, and policymakers, guiding strategic decision-making within the context of the Dillard dataset.

The implications of this analysis are substantial, catering to the needs of diverse stakeholders – from retail managers seeking operational optimizations to brand proprietors strategizing market presence, and policymakers aiming to foster a robust retail environment.

Data Cleaning, Data Processing & Exploratory Data Analysis

At the beginning of the project, we knew we needed a centralized way to access and run

queries to enable our team to operate more efficiently. Therefore, we chose to upload the datasets onto the SQL server provided by the MLDS program. We are given 5 datasets in total, so we name them accordingly and upload them to the server with access shared among our group. During this process, we need to initialize the primary key and set the data types accordingly.

Loading in big datasets like TRNSACT takes a long time and even the query checking the uploaded database takes more than a minute to execute. From there, we have a sense of the scale of our data. Then we are able to read data into pandas. The first step is to clean the data. We removed the last columns from all tables, which do not exist in the diagram. Then we check for missing values and there's none.

We start our EDA by making some plots. The first one is a pie chart based on the state information provided in the STRINFO table. It turns out that the top 2 states that have the most stores are TX and FL. From the TRNSACT table, we build visualizations that depict the difference between the original price and the sale price over time, showing a clear difference between them. We are also interested in the distributions of original and sale prices, and both show the majority less than $50.

From there, we had a good understanding of our datasets and then we merged them based on the identifier linked between each other. We first create a descriptive table to get a sense of each column and a heatmap to see the correlations. We also create a line graph containing profit and cost over time. These two lines look parallel to each other over time. Other line plots indicate that the selling quantity and discount rate are both strictly increasing over time.

Based on our business question, we build two additional plots describing high/low-profit counts by top brands and top states, both ranked by profit. We observe an overwhelmingly large count of high profit dominating the low profit in the graph by top brands. However, on the state level, the counts of high profit versus low profit are roughly the same.

Feature Engineering

We've created a new column named ‘High_Profit’ and dropped the ‘PROFIT’ column. Profits exceeding 100 are considered high profit and are coded as 1; otherwise, they are stored as 0. We chose 100 because this number balances the data. This High_Profit variable is our target variable. Additionally, we've generated the “discount_rate” column, calculated by subtracting the ‘SPRICE’ (sale price) from the ‘ORGPRICE’ (original price) and dividing by the ‘ORGPRICE’. Finally, we decided to include the variables ‘STATE’, ‘STORE’, ‘BRAND’, ‘Year’, ‘SPRICE’, ‘QUANTITY’, ‘ORGPRICE’, ‘COST’, ‘RETAIL’, and ‘discount_rate’ since they are related to brand, while the other variables are related to items. We also factorized the ‘STATE’ and ‘BRAND’ variables since they are categorical variables.

Modeling & Evaluation

After feature engineering all the variables, we are ready to fit the model. We fitted three different models, logistic regression, decision tree classifier, and k-neighbors classifier to compare the models’ performance and decide which model to use. Our response variable is ‘High_Profit’ and our predictor variables are the rest (stated in the feature engineering part). We first split the entire dataset into x (predictor variables) and y (response variable). Then, we divided x and y into X_train, X_test, y_train, and y_test by allocating 20% for the test set and 80% for the training set.

The first model we selected is logistic regression, implemented from the sklearn package. We fitted the model using X_train and y_train. For the logistic regression model, we added the L1 penalty term, which is the Lasso regularization. This regularization method can shrink less important features towards zero, especially beneficial with high-dimensional datasets. This can also avoid overfitting. And we used the ‘liblinear solver’. Upon predicting X_test, we evaluated the model's performance by comparing y_pred and y_test. We found out the AUC of testing data is about 0.9097 and the accuracy rate is 0.91. From the confusion matrix, the true negatives (17437) and true positives (13192) values are high

while the false negatives (344) and false positives (1386) are pretty small, which indicate a strong overall performance in correctly predicting both absence and presence of the target class.

The second model we selected is the decision tree classifier, which is also implemented from the sklearn package. Follow the same steps to train the model. We selected max_depth as 3 for this model. Finally, we got the AUC of around 0.9500 and the accuracy rate of 0.95. The true negatives (16344) and true positives (15439) are also high. The number of false positives (1386) and false negatives (344) are low. We also visualized the decision tree to observe its splits. At the first layer, the left side is determined by SPRICE being less than or equal to 167.985, while the right side is determined by discount_rate less than 0.488. Moving to the second layer, the left branch further splits into two sub-branches: SPRICE less than or equal to 146.995 on the left and cost less than or equal to 86.69 on the right. For the right branch, the split occurs between SPRICE less than or equal to 287.84 and SPRICE greater than 738.555.

The third model we explored is the K-neighbors classifier. Following the same steps as with the logistic regression model, we selected 5 as our number of neighbors. Subsequently, we achieved an AUC score of approximately 0.9491 for the testing set, with an accuracy rate of 0.95. The true negatives (16793) and true positives (15009) are also high. The number of false positives (937) and false negatives (774) are low.

After comparing these models, we identified the decision tree as our best model due to its highest AUC and accuracy rate when the maximum depth is increased. For instance, when adjusting the max_depth to 10, we achieved an accuracy of around 0.99 and an AUC of approximately 0.9912. This resulted in 17572 true negatives and 15646 true positives, with 158 false positives and 137 false negatives. This model performs the best compared to other models.

ROI Analysis

We want to evaluate the profitability and efficiency of an investment. It is commonly expressed as a percentage and is calculated using the following formula: (Retail Gain - Cost of Investment) / (Cost of investment). We observe that about 47.26% of products in Dillards had high profit grouping by state,

store, brand, and year, but above the average of the products (52.74%) are sold in low profit (even have negative value). We can see the average difference is 10.55 dollars. After fitting the model (random forest) with high true accuracy 0.95. The calculated ROI is approximately 12.56%. This means that for every dollar invested in the model, there is a return of 12.56 cents in profit. A positive ROI indicates that the model is generating more revenue than the total cost of investment. The model is showing a positive ROI, which is a promising sign. Stakeholders can expect a return of approximately 12.56% on their investment.

Conclusion

In conclusion, this study provides a detailed and comprehensive analysis of the factors influencing profitability in the Dillard's dataset. Utilizing analytical techniques such as logistic regression, decision tree classifiers, and k-neighbors classifiers, we have understood the effects of variables like store names, locations, sell prices, and discount rates in determining high and low profitability.

Our approach, which included extensive data cleaning, processing, exploratory data analysis, and feature engineering, enabled us to derive meaningful insights and patterns. The models' predictive accuracy and the decision tree model's superior performance illustrate the effectiveness of our analytical methods.

The ROI analysis further underscores the practical value of our findings. With a positive ROI of approximately 12.56%, this model stands as a testament to the potential financial gains that can be achieved through data-driven decision-making. Overall, through the insightful analysis of the Dillard dataset, we were able to provide a comprehensive approach to understanding and predicting retail profitability. By successfully identifying the key factors influencing high and low profits across different brands, stores, states, and over various years, our analysis offers valuable guidance for future strategies aimed at enhancing profitability in the retail sector.