Optimizing Dillard's Expansion: Using Sales Data and Consulting-Backed Estimation to Select the Next Local Store Location(s)

MLDS 400 Group 5
Xiyi Lin, Omar Shatrat, Fanqi Song, Tianyu Wu

Table of Contents

Executive Summary
Introduction
Data Cleansing
Exploratory Data Analysis (EDA)
Feature Selection and Engineering
Model Development and Evaluation
ROI Analysis
Limitation
Reference
Appendix

Executive Summary

In the ever-evolving landscape of retail expansion, strategic decision-making regarding new store locations plays a pivotal role in achieving sustained growth and maintaining market competitiveness. This paper addresses a pertinent business challenge faced by Dillard's. To measure overall success and profitability, we introduce key performance indicators (KPIs) such as *Average Cost of Goods Sold (COGS) per Item Sold*, *Month-over-Month Growth*, and *Return Percentage* for each store, derived from the sales data. Leveraging advanced methods including Logistic Regression, Decision Tree, Random Forest, SVM, and k-NN, our analysis provides actionable insights. These insights facilitate informed decision-making for optimal new store placement, categorized as either 'Successful' (1) or 'Not Successful' (0). Among those methods, the Random Forest model turns out to have a robust performance with a good weighted average efficacy, with 91% accuracy, 80% macro precision, 85% macro recall, and 82% macro F-1 score. This suggests the model's potential as a reliable predictor for future analyses.

Introduction

In the dynamic landscape of business operations, optimal resource allocation is imperative for sustained success. Navigating the rich landscape of 'skstinfo', 'strinfo', and 'trnsact' datasets, we embark on an analytical journey steeped in predictive modeling, which lies in training classification models to forecast the prospective success of new store locations.

Spanning the sales data, our goal is to conduct feature selection and engineering to better classify successful/unsuccessful stores, and further predict future store opening location(s) by using consulting-backed estimates of those features. This approach provides a definitive and measurable metric for the success of our predictive modeling, bridging the theoretical and the practical. As we delve into this analysis following the pipeline in Figure 1, the objective is not only to uncover insights but also to provide actionable strategies for informed decision-making in the ever-evolving retail realm.

Data Cleansing

Data Preparation and Unification

The initial phase of our data analysis involved the integration of the 'trnsact', 'skstinfo', and 'strinfo' tables into a unified data frame, as shown in Figure 2. This consolidation ensures a comprehensive view of transactional, departmental, and store information, laying the groundwork for subsequent analytical endeavors.

Data Integrity Measures

To safeguard the integrity of our dataset, rigorous actions were taken to handle null values and eliminate duplicates. These steps were crucial in mitigating potential biases and inaccuracies that could influence subsequent analyses. Additionally, the 'SALEDATE' column transformed into DateTime format 'YYYY-MM', enabling nuanced temporal analyses.

Specifically, special attention was given to addressing missing values, focusing primarily on the 'Cost', 'Retail Price', 'Sale Price', and 'AMT' columns. The 'AMT' column was imputed using an obvious mathematical relationship with the 'Sale Price' and 'Quantity' columns. The formula utilized for imputation was as follows: AMT = Sale Price * Quantity. For the 'Cost' column,

missing values were imputed by calculating the average cost within specific groups identified by the Stock Keeping Unit, i.e., 'SKU'. This method ensures that missing 'Cost' values are replaced by the average cost of the corresponding SKU, providing a representative estimate.

Data Enrichment: Geo-encoding
As part of the data enrichment process, geo-encoding was performed by leveraging information from the 'CITY' and 'STATE' columns to generate corresponding 'Latitude' and 'Longitude' columns. This geo-encoding initiative aims to enhance the dataset with spatial information, enabling geographical analyses and visualizations. The process involved mapping city and state information to geographical coordinates, providing a more comprehensive understanding of the dataset's geographical distribution. This geo-encoded data can be instrumental in uncovering location-based insights and trends within the context of the broader analysis.

Exploratory Data Analysis (EDA)

Store Distribution Insights
Our exploration begins with an in-depth analysis of the 'strinfo' table, unraveling insights into the distribution of stores across diverse states and cities (Figure 6). Predominantly, Texas, Florida, Arkansas, Arizona, and Ohio emerge as retail powerhouses, boasting the highest concentrations of stores, as shown in Figure 3. Moreover, Figure 4 shows that cities such as Little Rock, Gilbert, Olathe, San Antonio, and Houston stand out as strategic focal points, potentially influenced by factors like heightened consumer demand and meticulous strategic planning.

Cost and Retail Price Dynamics
Delving into summary statistics for cost and retail prices unveils a well-balanced distribution in Figure 5, with the majority of products exhibiting moderate cost and retail price values. Noteworthy outliers, representing high-cost and high-retail price items, may signify premium products within the retail landscape. A robust positive correlation (0.896) between cost and retail prices indicates a general trend: higher product costs correlate with elevated retail prices.

Insights into Profit Margins
Profit margin, a pivotal metric delineating the disparity between retail price and cost, takes center stage in our analysis. Calculated as (Retail Price - Cost) / Cost, this metric unveils profound insights into the financial health of businesses. Our identification of the top five states with the highest profit margins—Arkansas, Oklahoma, Ohio, Texas, and Tennessee—signals not only a significant store presence but also efficient cost management and well-executed pricing strategies. This observation underscores these states' robust financial performance and strategic acumen.

Feature Selection and Engineering

Key Performance Indicators (KPIs)
We implemented computational codes to derive essential Key Performance Indicators, i.e., KPIs (A'alona, 2020) which are critical in assessing the performance and success of the dataset, as seen in Table 2's demo:

1. **Average Cost of Goods Sold (COGS) per Item Sold**: Representing the average cost incurred to produce or purchase the goods that were sold within a given period. It's calculated by dividing the total Cost of Goods Sold by the number of items sold. The interpretation of this metric can provide insights into the efficiency of the business operations and profitability. Specifically, a *lower* average COGS per item sold indicates that we are generating sales while keeping the cost of producing those goods relatively low.

2. **Month-over-Month (MoM) Growth**: Indicating the percentage change in gross profit[1] from the previous month. A positive MoM Gross Profit Growth indicates an increase in gross profit, while a negative value indicates a decrease.

$$\text{MoM Gross Profit Growth} = \left(\frac{\text{Gross Profit in Current Month} - \text{Gross Profit in Previous Month}}{\text{Gross Profit in Previous Month}}\right) \times 100$$

3. **Return Percentage**: Expressing the proportion of sales revenue returned to the business as product returns, a higher percentage implies a larger share of sales being returned, impacting profitability and operational efficiency.

Outliers
During the analysis of store performance metrics, stores 503 and 3209 were identified as outliers due to their extreme values in MoM Growth. To enhance the reliability of our analysis and predictions, we decided to dismiss Stores 503 and 3209 from the dataset. This step aims to prevent skewed results and inaccuracies that may arise from the presence of outliers.

Exploration of Interaction Term's Impact
Understanding the distribution of individual features and potential patterns between 'success' categories based on the pair plot from Figure 8, we introduced an interaction term, *AM_Interaction*, by multiplying the *AvgCOGS_peritem* and *MoMGrowth* columns. This step was taken to account for the synergistic or multiplicative impact of the two variables. From Figure 9, we can see the perfect correlation between the interaction term and the MoMGrowth column. Hence, we decided to drop the *AvgCOGS_peritem* column, that is, the *Average Cost of Goods Sold (COGS) per Item Sold* feature, as well as that interaction term.

Rationale for Success Criteria
The determination of a store's success is multifaceted and revolves around key performance indicators. According to our selected features, a store is deemed successful if it meets the following criteria:

- **MoM Growth**:
- Rationale: A store is deemed successful if it is above the 50th percentile[2] in *MoM Growth*. This criterion underscores the importance of sustained growth, positioning stores within the high-spread interval.

[1] *Gross Profit*: Representing a store's profit after deducting the cost of goods sold from total revenue, higher gross profit signals enhanced profitability.
[2] We choose 50th percentile (median) to ensure a balanced split, where approximately half of the stores are classified as successful and the other half as not successful, and it is less sensitive to extreme values (outliers) than the mean.

- Implication: Placing a premium on stores with *MoM Growth* above the median signals a strategic emphasis on sustained and above-average growth. This aligns with our commitment to fostering dynamic and thriving retail outlets.

- **Return Percentage**:
- Rationale: Success is further contingent on the *Return Percentage* falling within the low-spread interval, specifically below the median. This dual criterion ensures that a successful store not only achieves growth but also maintains efficiency in cost management and customer satisfaction.
- Implication: Stores achieving success can effectively manage this challenge, maintaining a low return percentage. This highlights a commitment to customer satisfaction and product quality.

This threshold-based classification gives an appropriate separation with 15.3% of stores being successful, while the remaining 84.7% of the stores being unsuccessful, laying the foundation for addressing this classification problem (Figure 7, 11).

Model Development and Evaluation

To tackle the class imbalance, SMOTE was applied exclusively to the training set, oversampling the minority class to create a more balanced representation. The distribution of the target variable before and after SMOTE was examined as shown in Figure 10, emphasizing the effectiveness of the oversampling technique in mitigating the class imbalance.

Model Comparison and Selection

After successfully addressing the class imbalance through the application of SMOTE exclusively on the training set, our focus shifted to the implementation of various classification algorithms on the balanced data. Training data has been applied to several classification algorithms, including Logistic Regression, Decision Tree, Random Forest, SVM, and k-NN. Different algorithms produce various results, as shown in Table 1.

Table 1. Classification Model Metrics Comparison

Models	Metrics			
	Accuracy	Macro Precision	Macro Recall	Macro F-1 Score
Logistic Regression	0.76	0.64	0.77	0.65
Decision Tree	0.89	0.78	0.75	0.76
Random Forest	0.91	0.80	0.85	0.82
SVM	0.65	0.56	0.61	0.54
K-NN	0.70	0.62	0.73	0.60

Recommendation - Decision Region
The result above implies that the Random Forest model is recommended to select the promising location(s) for Dillard's next local store. Further, based on consulting-backed estimations of selected features (Table 3), our model has identified COLUMBIA and PHOENIX as the most promising locations among the five under consideration. The rationale behind selecting these two areas is apparent: they exhibit above-median Month-over-Month (MoM) Growth, below-median Return Percentage figures, and have a track record of successful store openings in the training set. Their prominence signifies strong potential and suitability based on the parameters and criteria integrated into our model, indicating them as prime choices for further exploration and investment considerations.

ROI Analysis

The US clothing market is estimated to be valued at $351.4 billion, with a Compound Annual Growth Rate (CAGR) of 1.93%. The projected annual market size is expected to reach $358.18 billion. In 2022, Dillard's, a popular clothing department store, reported a revenue of $6.9 billion, operating expenses (OpEx) of $1.67 billion, and operates 277 stores. The revenue per Dillard's store location is approximately $24.91 million, with operating expenses at $6.04 million, resulting in a margin of $18.87 million per location.

Should Dillard's expand with the market, targeting a 5.35% growth rate, our model gives insights on what type of return they can expect. The fixed opening costs for each new store are $100,000, and the typical lease term is 24 months with a rent of $61.40 per square foot. The average size of a Dillard's store is 250,000 square feet, leading to a monthly lease cost of $15.35 million. With these figures, the annual profit per store is projected to be $3,416,425.99. The success rate for new stores based on our model is approximately 15.3%, resulting in an expected 0.82 successful stores in the coming year. Consequently, the Return on Investment (ROI) is calculated at 2,794,476.92.

Limitation

The efficacy of our predictive modeling heavily depends on the quality and accuracy of available datasets, particularly 'skstinfo', 'strinfo', and 'trnsact'; any limitations or biases within these datasets may compromise the robustness of our analyses and subsequent predictions. Additionally, our models' predictive power relies on the assumption that historical trends will persist in an ever-evolving market landscape, and unforeseen shifts in consumer behavior, economic conditions, or competitive dynamics could challenge the accuracy of our predictions, such as regulatory changes or unforeseen market events, introducing an element of uncertainty.

Reference

Jason A'alona, 2020. Podium. (n.d.). Top KPIs for Retailers. Retrieved from https://www.podium.com/article/top-kpis-for-retailers/

Appendix

Table Part

Table 2. Demo of KPI Values

	STORE	AvgCOGS_peritem	MoMGrowth	ReturnPercentage	CITY
0	102	18.656348	0.102883	7.297510	TAMPA
1	103	17.761179	-1.857504	8.766210	ST LOUIS
2	107	16.975893	0.350877	8.382016	HURST
3	202	15.533738	0.190422	8.271629	TAMPA
4	203	16.687797	-0.614171	10.119624	CHESTERFIELD
...	...	...	...	...	...
326	9709	13.221052	-1.186029	6.835200	GREELEY
327	9804	14.891369	0.068109	7.260616	LAWTON
328	9806	17.679472	0.040775	4.250364	MABELVALE
329	9906	5.600000	NaN	0.000000	LITTLE ROCK
330	9909	13.783600	-0.668611	5.773841	CHEYENNE

Table 3. Consulting-Backed Estimation Data Used for Predicting Stores of Interest's Success

	MoMGrowth	ReturnPercentage	Latitude	Longitude	CITY
0	20.401200	0.00200	35.395	-95.814	OKLAHOMA CITY
1	0.812300	4.28038	39.204	-76.690	COLUMBIA
2	-0.125950	10.29030	32.380	-86.312	MONTGOMERY
3	0.067734	6.32000	33.415	-111.835	MESA
4	0.472900	5.20390	33.451	-112.016	PHOENIX

Picture Part

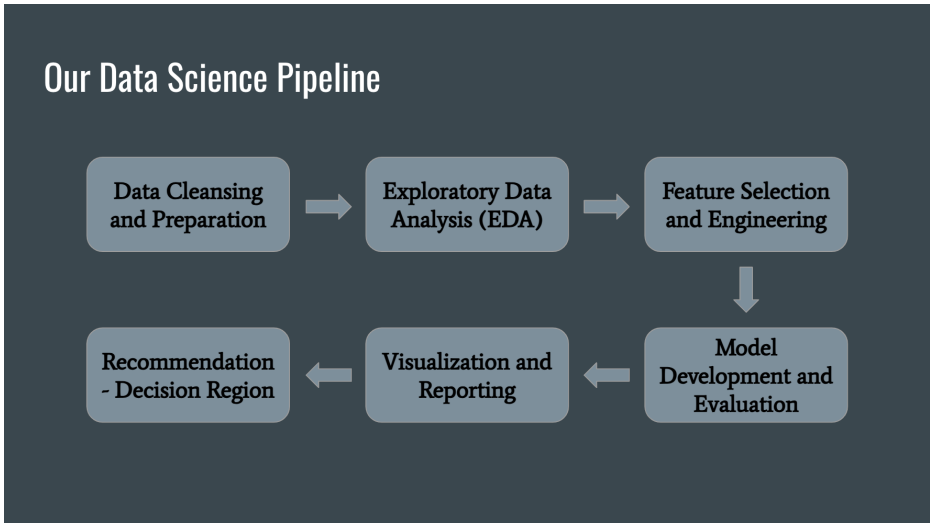

Figure 1. Our Data Science Pipeline

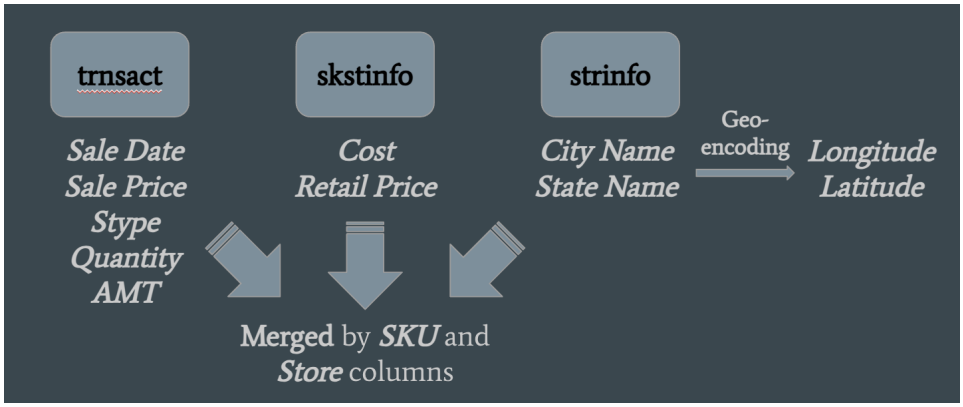

Figure 2. Data Preparation

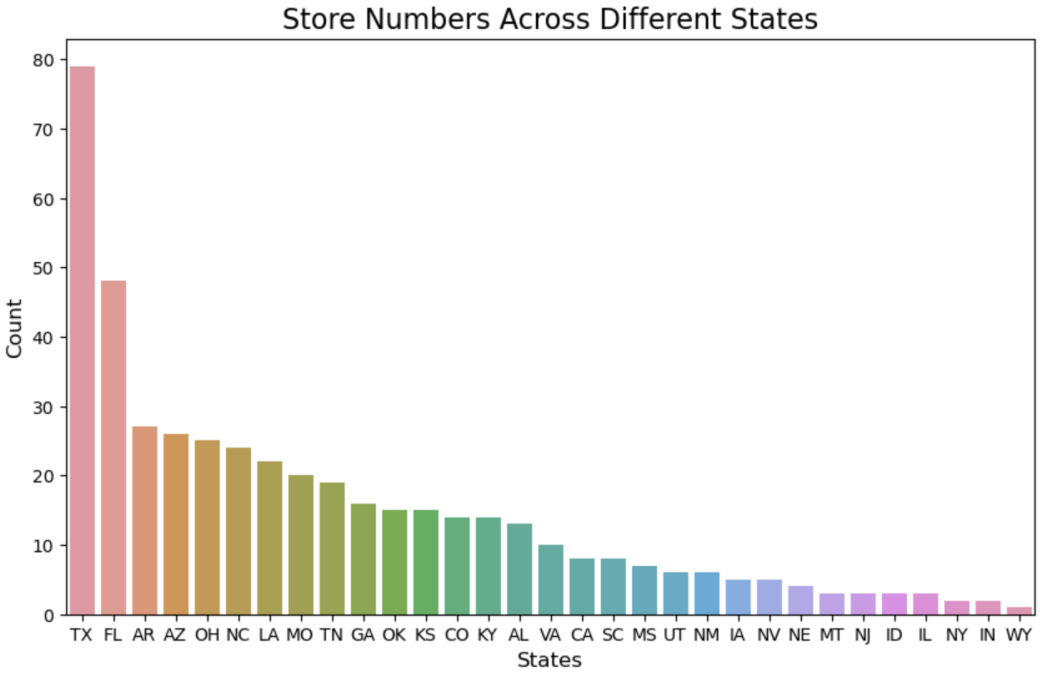

Figure 3. Store Numbers Across Different States

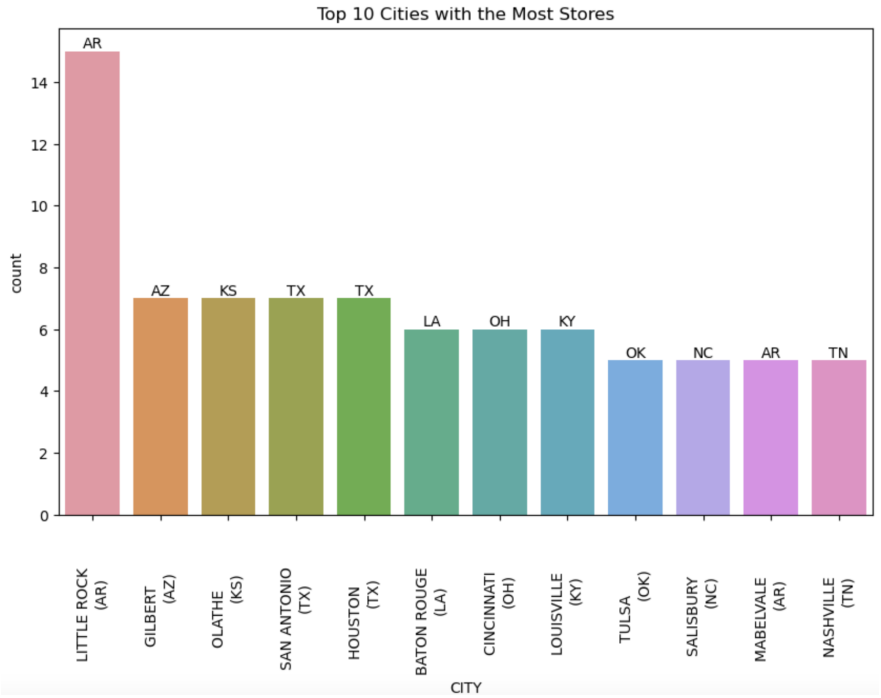

Figure 4. Top 10 Cities with the Most Stores

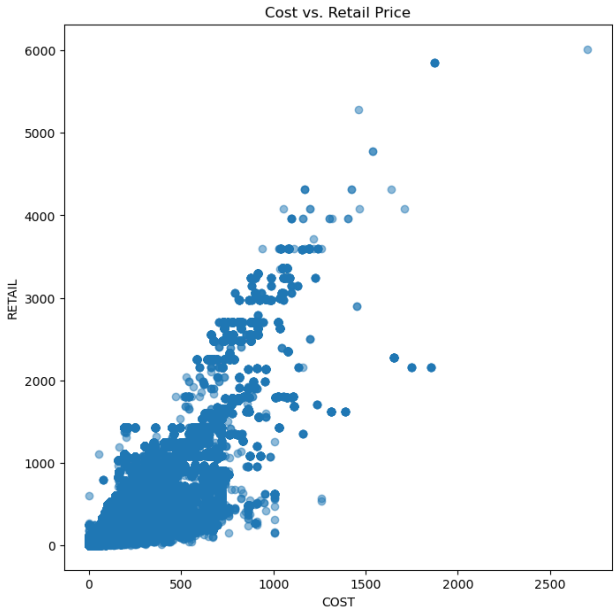

Figure 5. Cost vs. Retail Price

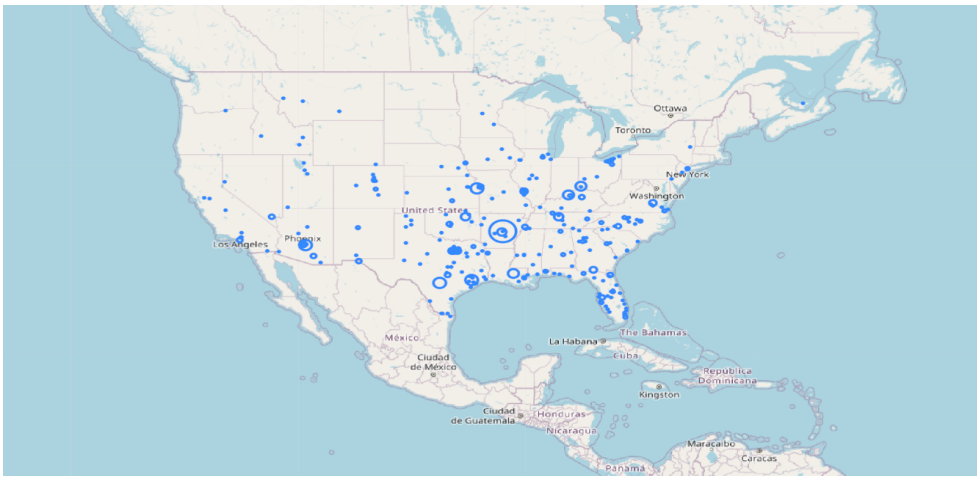

Figure 6. Store Distribution on the Map

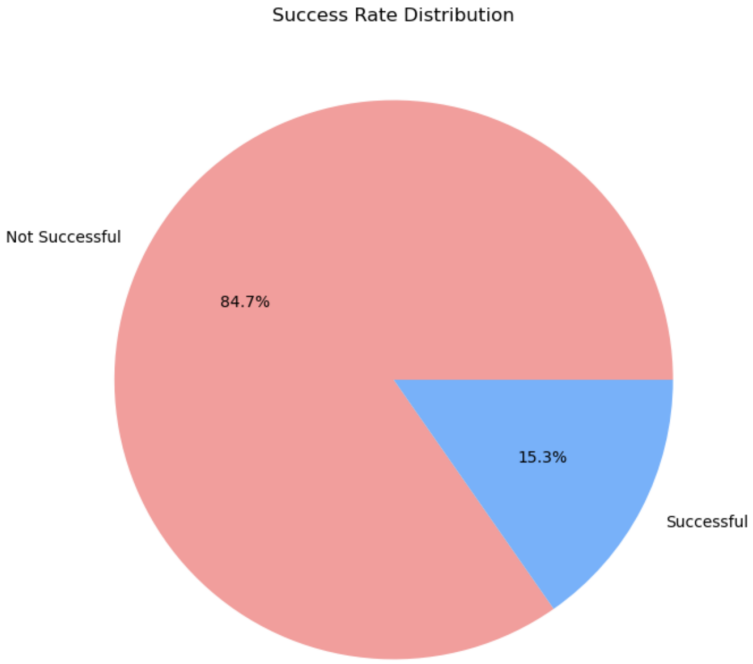

Figure 7. Store Success Rate Distribution

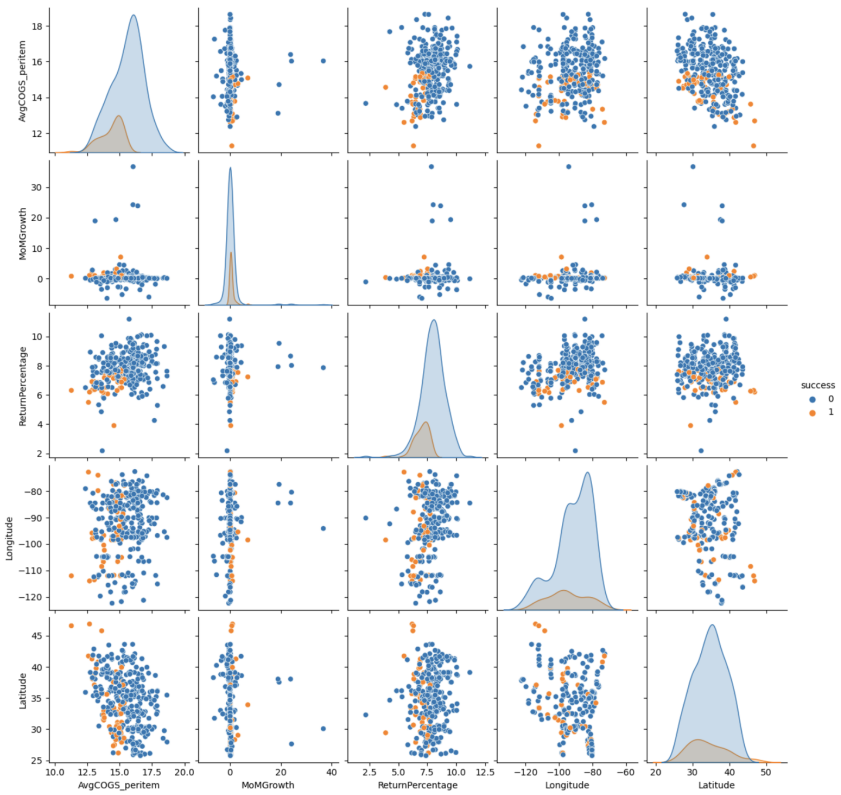

Figure 8. Pairplot of newly created features (KPI)

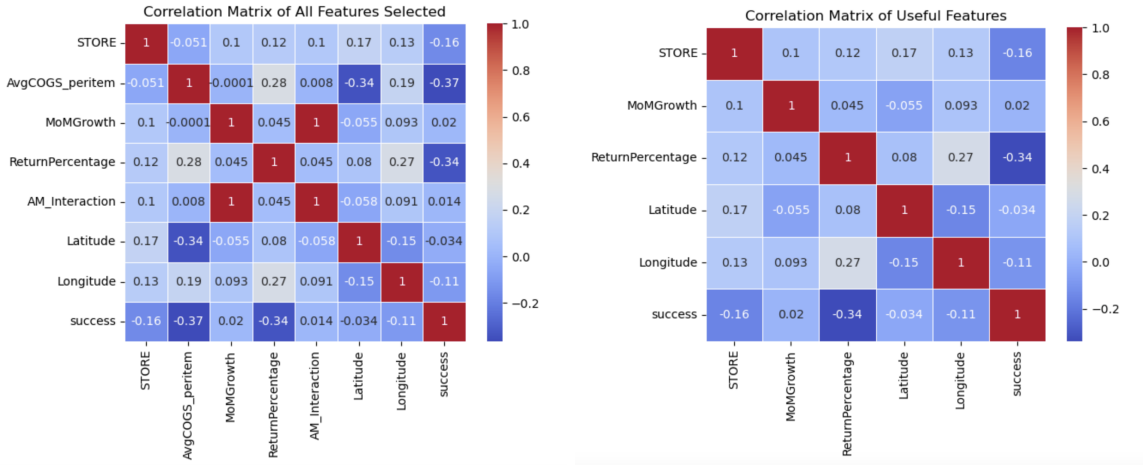

Figure 9. Correlation Matrix (Before vs. After)

```
Class distribution before SMOTE:
success
0    220
1     41
Name: count, dtype: int64

Class distribution after SMOTE:
success
0    220
1    220
```

Figure 10. SMOTE on the Imbalanced Dataset

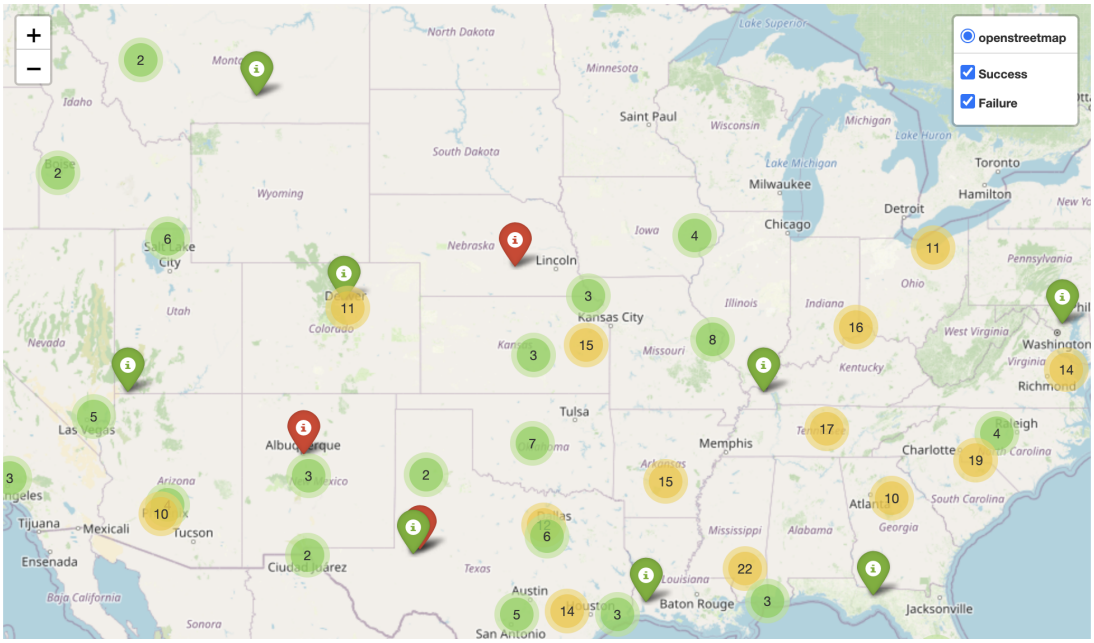

Figure 11. Demo of Store Success Distribution on the Map