MLDS 400 Project Report

Yumin Zhang, Wei Wang, Jiayue Tian, Ye Joon Han

Table of Contents

Executive Summary

This project serves Dillard's, one of America's largest fashion retailers with a storied history dating back to 1938 and achieving $6.34 billion in revenue in 2020. Facing the challenge of product returns, Dillard's aimed to leverage historical sales data to predict the likelihood of returns, thereby enabling more efficient return management and inventory control. Our analysis showed that 44.7% of Dillard's products sold were discounted, a strategy that did not significantly contribute to profit gains. Furthermore, extended discount periods were associated with a 28% decrease in long-term benefits, emphasizing the need for a strategic approach to discounting. By utilizing a Random Forest model, augmented with SMOTE to address class imbalance, we achieved a reliable predictive capability, indicated by a true positive rate (TPR) of 1 and a false positive rate (FPR) of 0.2. The Return on Investment (ROI) analysis underscored the financial viability of implementing the predictive model, with a 15% ROI gain at a sensitivity threshold of 0.57. The model's deployment, costing $963,400, is projected to yield a retail gain of over $1.1 million for Dillard's, validating the investment in our data-driven solution. This strategic tool empowers Dillard's to optimize pricing strategies, reduce unnecessary losses from returns, and enhance overall operational efficiency (see Figure 1 in the Appendix for ROI analysis).

Introduction

Dillard's, an established leader in the American fashion retail industry since 1938, faces the intricate challenge of product returns—a common yet complex issue that significantly impacts revenue and operational efficiency. With 2020 revenues soaring to $6.34 billion, the company recognizes the potential of leveraging big data to enhance its return management processes. This report details our endeavor to utilize Dillard's extensive historical sales data to predict the likelihood of product returns, aiming to bolster decision-making and refine customer engagement strategies. The foundation of our analysis lies in a comprehensive examination of Dillard's transactional datasets, intricately connected within a sophisticated database structure. Our journey began with the meticulous extraction of data from multiple datasets, including STRINFO, SKSTINFO, and TRNSACT, facilitated by the MLDS PostgreSQL cloud server. The information extracted—ranging from SKU details to transactional histories—was pivotal in constructing a cohesive framework for our predictive analysis. (Refer to Figure 2 in the Appendix for a visualization of sales trends).

Data Preparation Methodology

The data preparation stage is crucial in the analytics lifecycle, especially for a data-rich environment like Dillard's, where strategic decisions hinge on the integrity and granularity of data insights. This project embarked on a rigorous data preparation methodology, ensuring that the foundation for subsequent analysis was both robust and reflective of the real-world complexities of retail operations.

Data Acquisition: Our initial step involved harnessing Dillard's vast transactional data, housed within their PostgreSQL cloud server. Through precise SQL queries, we targeted essential features that could provide insight into sales trends, product performance, and customer purchasing behavior. This data was then imported into Python, where Pandas DataFrames served as the backbone for our manipulation and analysis.

Data Cleaning: Upon acquisition, the data underwent an extensive cleaning process. We addressed missing values, outliers, and duplicates that could skew our analysis, ensuring a dataset that accurately represented true business operations. Columns with ambiguous or redundant information were either clarified or removed, streamlining the dataset to the most impactful variables.

Handling Class Imbalance with SMOTE: After the thorough data cleaning process, we confronted a common challenge in predictive modeling, especially in retail analytics: class imbalance. This occurs when the occurrences of a class of interest (in our case, 'product returns') are significantly lower compared to the other class. To address this, we employed the Synthetic Minority Over-sampling Technique (SMOTE). This approach helped us to balance our dataset effectively, enhancing the predictive performance of our models. By applying SMOTE, we ensured that our models do not exhibit a bias towards the majority class and can generalize well when predicting product returns, a critical aspect of our analysis for Dillard's.

Feature Engineering: To enrich our analysis, we engineered new features that could shed light on underlying patterns. For instance, we calculated the markup percentage as a measure of profitability and created aggregated transaction counts to assess sales volume. These features were intended to provide a multifaceted view of the data, going beyond the surface-level numbers.

Data Transformation: Numerical columns underwent a log transformation to normalize their distribution, which is critical for the performance of many predictive models. This step was particularly important for handling skewed data, such as transaction amounts and item pricing, which could lead to biased insights if left untransformed (see Figures 3 and 4 in the Appendix).

Exploratory Data Analysis (EDA): With a clean and enriched dataset, we embarked on an exploratory data analysis. We delved into the discount strategies by calculating the total discounted items per month and their correlation with total sales. Additionally, we examined the retail price versus the original price to understand the revenue impact over time.

Conclusion: The preparatory work done in this phase laid a strong analytical foundation for Dillard's. By meticulously refining the dataset, we ensured that our predictive models and analyses would be both accurate and actionable, thereby empowering Dillard's with the tools to make informed business decisions.

Model Building and Evaluation

Model Building: With the data meticulously prepared, we transitioned into the model building phase. Our approach was methodical, employing a combination of logistic regression and random forest models to harness different aspects of the data's predictive power. Logistic regression was chosen for its interpretability and efficiency, providing a clear view of the relationship between features and the likelihood of product returns. The random forest model, known for its high accuracy and robustness to overfitting, allowed us to capture more complex patterns and interactions between variables. Each model was carefully tuned to balance the trade-offs between sensitivity and specificity. We iterated through various hyperparameters, utilizing cross-validation to avoid overfitting and ensure that the models generalized well to unseen data. The final models were selected based on their performance on a validation set, which was a hold-out portion of the dataset not used during the training phase.

Model Evaluation: The models' performance was assessed using a confusion matrix, which provided insights into true positives, false positives, false negatives, and true negatives. For a more nuanced view, we calculated additional metrics such as precision, recall, F1-score. Our logistic regression model achieved an accuracy of almost 0.579 and an F1-score of 0.715. The random forest model, with hyperparameters fine-tuned to the data, achieved an accuracy of 0.799 and F1-score of 0.885 indicating a high level of predictive capability. To translate model performance into business impact, we performed a cost-benefit analysis using the models' probability thresholds. This analysis helped us understand the potential financial implications of each model's predictions, guiding us in selecting a threshold that maximized ROI while maintaining a prudent sensitivity level.

Results and Discussion

The results of our modeling efforts reveal compelling insights into the operational dynamics of Dillard's and provide a pathway to enhancing the efficiency of its retail processes.

Discount Strategy Insights: Our analysis indicates that 44.7% of products sold were discounted, a strategy that resulted in an unimpressive gain when considering the broader financial context. This finding suggests that while discounts may drive short-term sales, their overuse could be detrimental to long-term profitability. The data revealed a stark potential for a 28% decline in long-term benefits when discounts are not strategically employed.

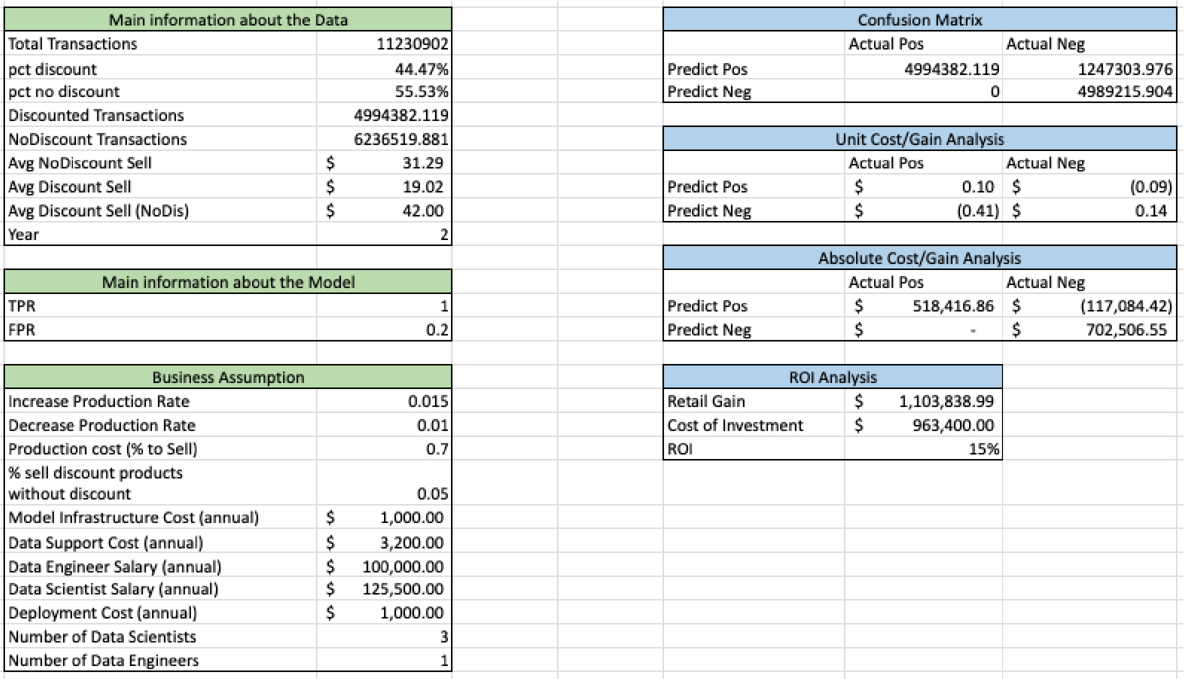
Model Performance and Financial Implications: The performance of the logistic regression and random forest models was critically evaluated to ensure effectiveness and reliability. The logistic regression model achieved an accuracy of 0.579 and an F1-score of 0.715, reflecting its capability to predict returns with a reasonable balance of precision and recall. Although this model didn't achieve extremely high accuracy, its F1-score suggests a decent balance in terms of precision and recall, which is vital in the context of predicting product returns. In contrast, the random forest model displayed a stronger performance with an accuracy of 0.799 and a recall of 0.839. This indicates that the random forest model was more effective in correctly identifying instances of product returns, which is crucial for Dillard's to manage inventory and minimize losses due to returns. The higher recall rate is particularly significant in this business context, where identifying potential returns accurately can lead to more effective strategies to reduce them. The financial analysis, leveraging the outputs from these models, highlighted a projected 15% ROI gain, showcasing the financial viability of integrating these predictive models into Dillard's decision-making framework. With a sensitivity threshold of 0.57, these models strike a prudent balance, proficiently identifying potential returns while minimizing false positives. This balance is essential for Dillard's, as it ensures that the strategies derived from these models are both cost-effective and impactful, enabling the company to make informed decisions that can lead to improved profitability and enhanced customer satisfaction.

Discussion of Findings: The ROI and sensitivity metrics suggest that the implementation of predictive models can significantly enhance Dillard's inventory management and discounting practices. The models offer a granular view of transactional patterns, enabling the retailer to identify which discount strategies might lead to increased returns and diminished value over time. However, the models also bring to light the importance of strategic discounting. By refining their discount strategies, Dillard's could potentially avoid the 28% decline in product value associated with protracted discount periods. Furthermore, the insights could aid in optimizing stock levels to reduce overstocking or understocking, particularly for items with a higher likelihood of return.

Conclusion: The results of this project serve as a testament to the transformative potential of data science in retail operations. For Dillard's, the integration of predictive analytics into their business strategy could not only minimize the financial impact of returns but also fortify customer satisfaction and loyalty through improved product offerings and pricing strategies.

Appendix

Figure 1. ROI Analysis



Main information about the Data	
Total Transactions	11230902
pct discount	44.47%
pct no discount	55.53%
Discounted Transactions	4994382.119
NoDiscount Transactions	6236519.881
Avg NoDiscount Sell	$ 31.29
Avg Discount Sell	$ 19.02
Avg Discount Sell (NoDis)	$ 42.00
Year	2

Main information about the Model	
TPR	1
FPR	0.2

Business Assumption	
Increase Production Rate	0.015
Decrease Production Rate	0.01
Production cost (% to Sell)	0.7
% sell discount products without discount	0.05
Model Infrastructure Cost (annual)	$ 1,000.00
Data Support Cost (annual)	$ 3,200.00
Data Engineer Salary (annual)	$ 100,000.00
Data Scientist Salary (annual)	$ 125,500.00
Deployment Cost (annual)	$ 1,000.00
Number of Data Scientists	3
Number of Data Engineers	1

Confusion Matrix

	Actual Pos	Actual Neg
Predict Pos	4994382.119	1247303.976
Predict Neg	0	4989215.904

Unit Cost/Gain Analysis

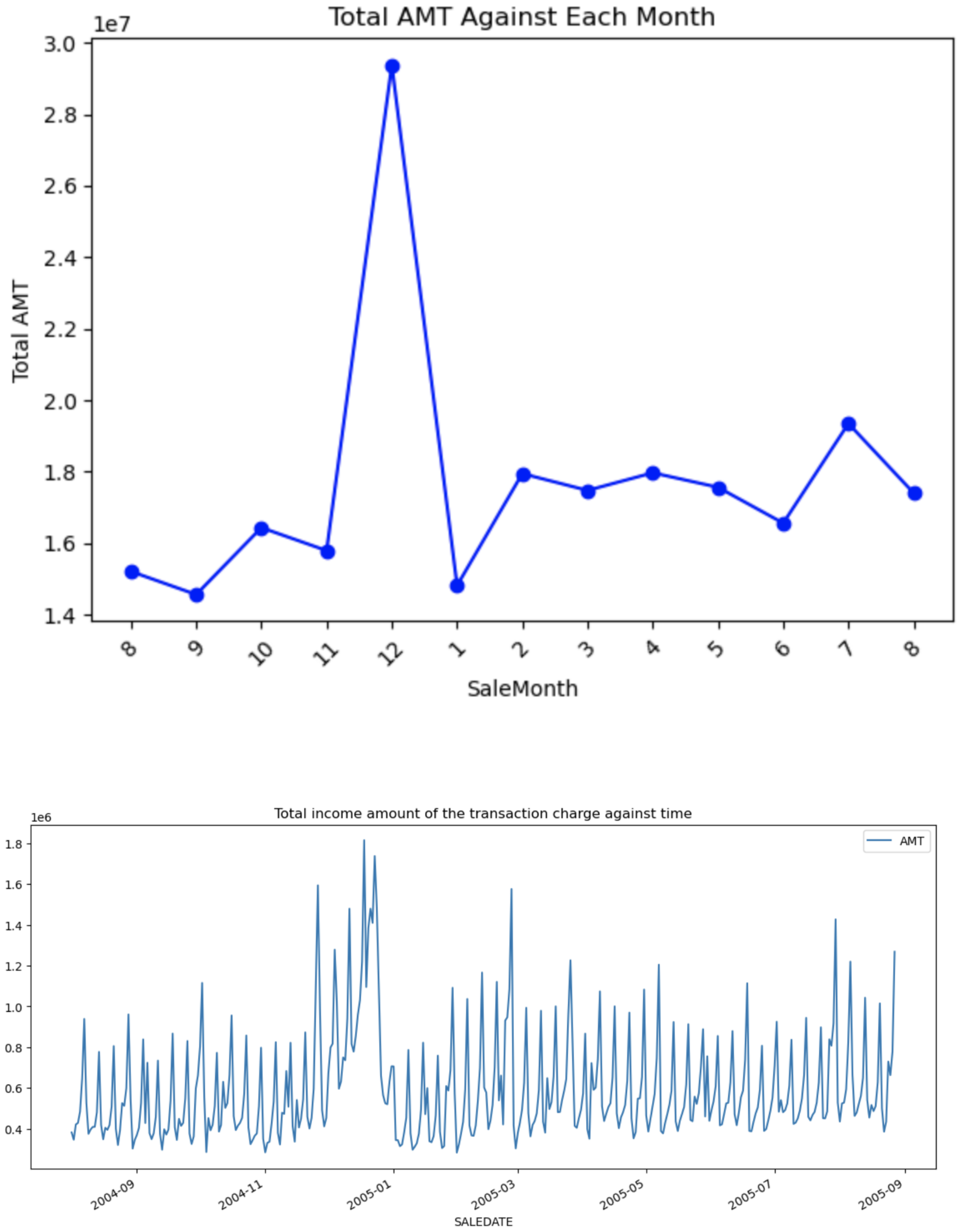
	Actual Pos	Actual Neg
Predict Pos	$ 0.10	$ (0.09)
Predict Neg	$ (0.41)	$ 0.14

Absolute Cost/Gain Analysis

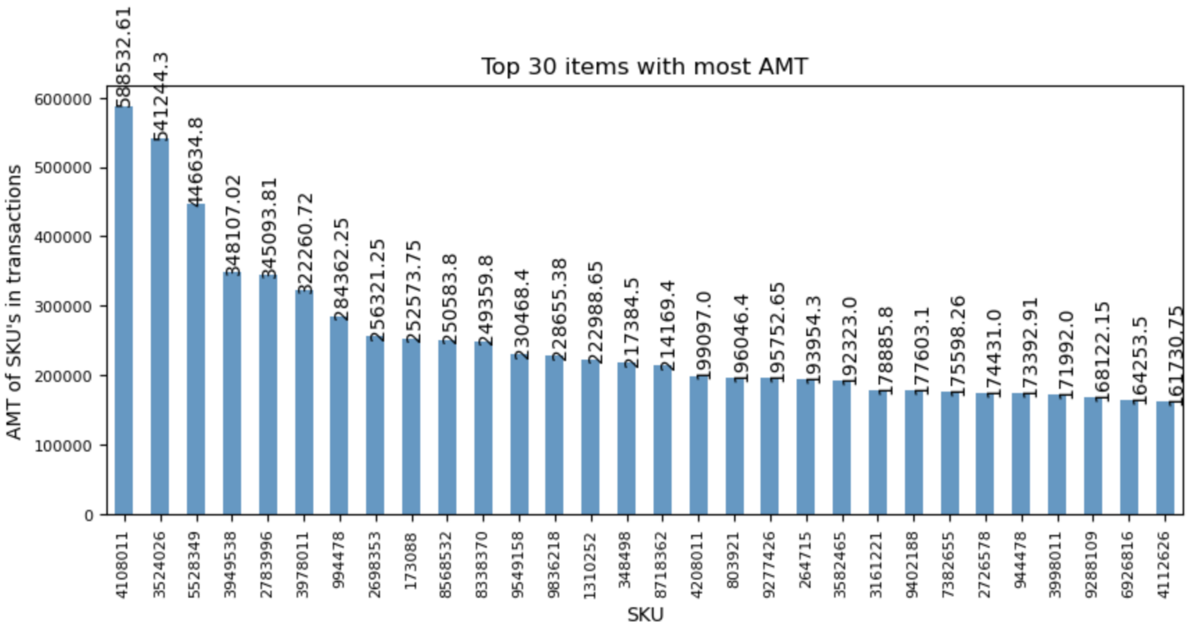
	Actual Pos	Actual Neg
Predict Pos	$ 518,416.86	$ (117,084.42)
Predict Neg	$ -	$ 702,506.55

ROI Analysis	
Retail Gain	$ 1,103,838.99
Cost of Investment	$ 963,400.00
ROI	15%

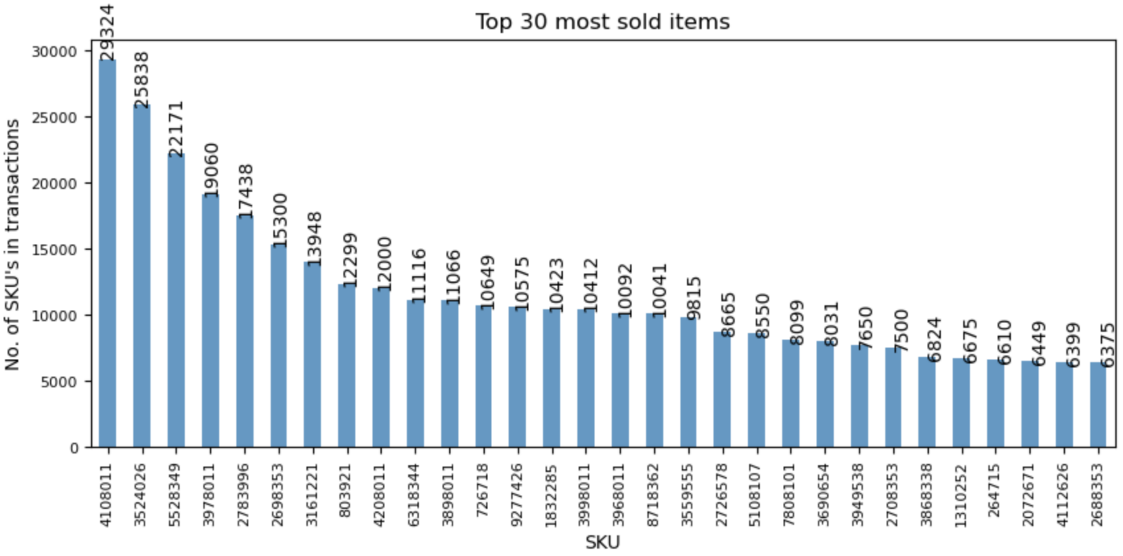
Figure 2


Total AMT Against Each Month
1e7
3.0
2.8
2.6
2.4
2.2
2.0
1.8
1.6
1.4
Total AMT
8
9
10
11
12
1
2
3
4
5
6
7
8
SaleMonth
Total income amount of the transaction charge against time
1e6
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
AMT
2004-09
2004-11
2005-01
2005-03
2005-05
2005-07
2005-09
SALEDATE

Top 30 items with most AMT



Top 30 most sold items



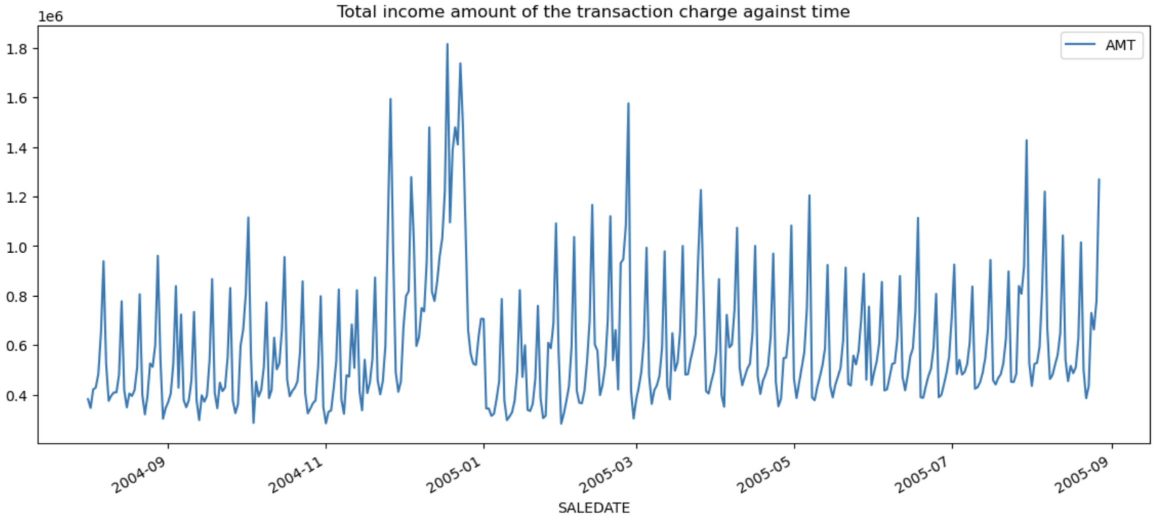

Total income amount of the transaction charge against time
1e6
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
AMT
2004-09
2004-11
2005-01
2005-03
2005-05
2005-07
2005-09
SALEDATE

Figure 3: Before Data Transformation

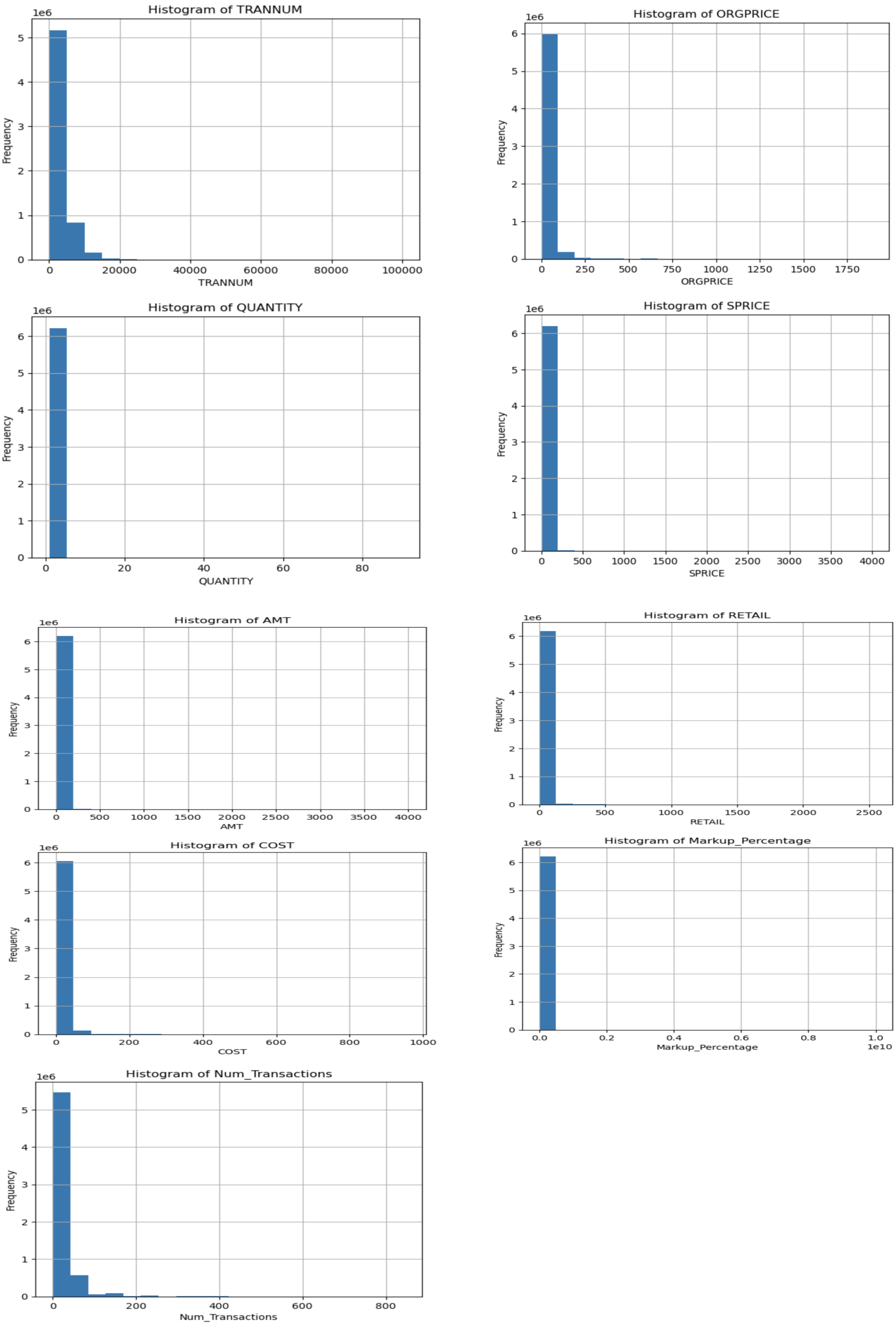
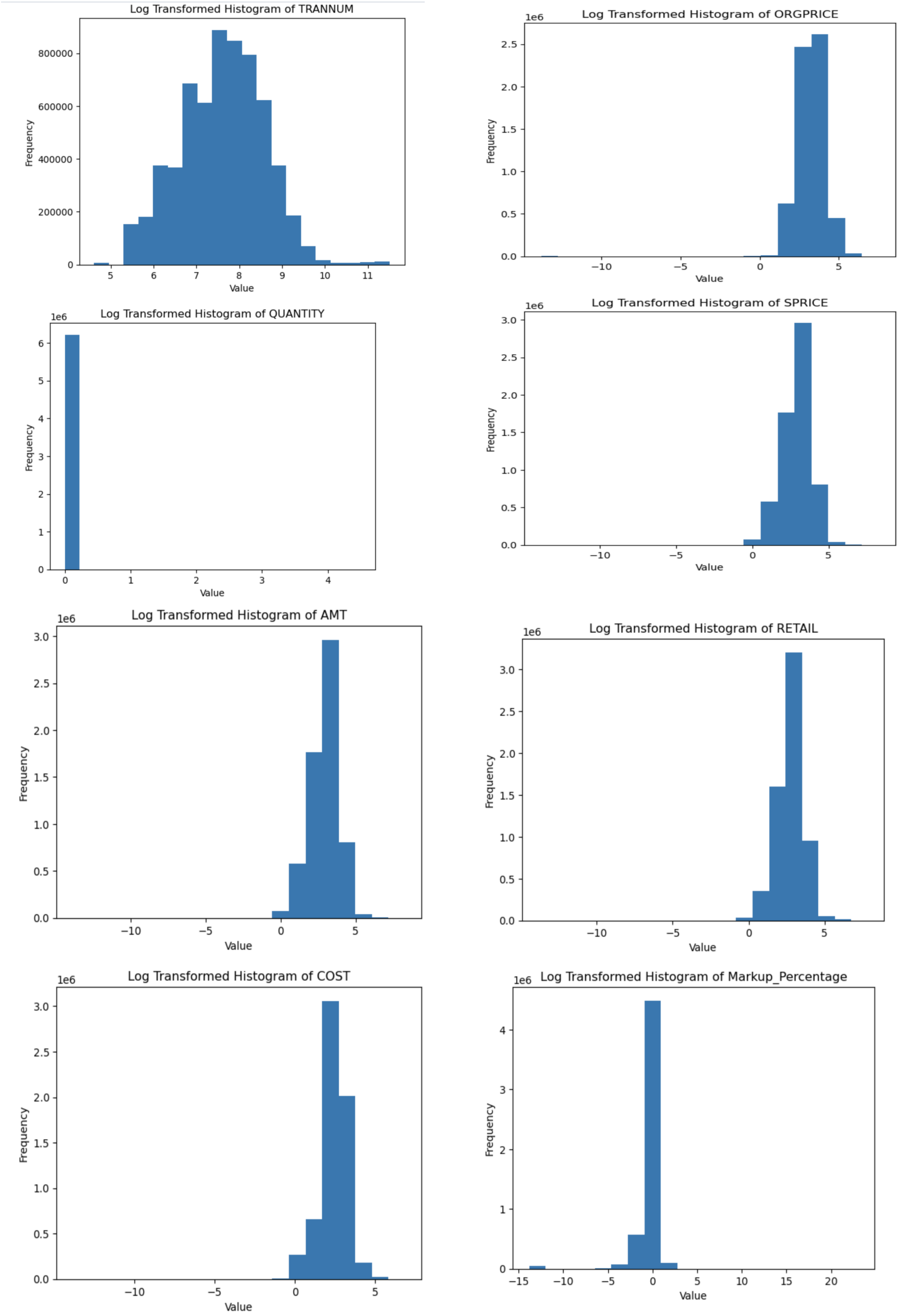


Figure 4: After Data Transformation


Log Transformed Histogram of TRANNUM
Frequency
Value
Log Transformed Histogram of ORGPRICE
1e6
Frequency
Value
Log Transformed Histogram of QUANTITY
1e6
Frequency
Value
Log Transformed Histogram of SPRICE
1e6
Frequency
Value
Log Transformed Histogram of AMT
1e6
Frequency
Value
Log Transformed Histogram of RETAIL
1e6
Frequency
Value
Log Transformed Histogram of COST
1e6
Frequency
Value
Log Transformed Histogram of Markup_Percentage
1e6
Frequency
Value