Project Summary: Dillard's Predictive Analysis for Inventory Cost Improvement

<u>1.Objective and Executive Summary</u>: Our project's core objective is to develop a predictive model tailored for forecasting inventory selling prices by projecting revenue by each quarter across diverse product categories with the goal to reduce costs from overstocking. By leveraging detailed transaction data, including purchase frequencies, store-level data on customer demographics, and SKU-level descriptive data on the brand and class of an item, we aim to align inventory levels with projected sales. This strategy is designed to reduce inventory-related costs, such as storage and depreciation, and mitigate the risks of overstocking, particularly for in-demand and seasonal items. Our model's effectiveness is highlighted by a significant quarter-over-quarter reduction in inventory costs, amounting to $117.9 million. This represents an improvement of 21.6% over the baseline model across three quarters for 2005.

<u>2.Data Description and EDA</u>

In the SKU info table, our analysis extends beyond basic attributes like brand and color. We examine how these attributes correlate with sales trends and inventory turnover rates. For instance, our findings indicate that the popularity of 'POLO FAS' significantly influences stocking decisions, aligning with customer preferences. In the STR Info table, the store distribution analysis includes not just geographical spread but also market penetration and sales performance per region. Our outlier removal in the transactions table was based on statistical criteria, such as z-scores and interquartile ranges, ensuring data integrity and relevance. In the skst Info table, null values were checked and statistical measures (mean, max, min) for cost and retail were assessed. Boxplots and histograms revealed both average retail and cost prices are right skewed with outliers. In the Dept info analysis, column data types were checked, revealing 60 unique departments with no missing data or empty spaces. A visualization confirmed each department, and its description appears once.

3.Feature Engineering

Data Preparation: We structured our data into a comprehensive data frame where each row captures a unique SKU in each store. This data frame was constructed by grouping each transaction table by the Stock Keeping Unit (SKU) and the location of each store. This grouping enabled us to streamline our dataset, reducing both the length of the data frame and computational runtime, along with setting the foundation for feature creation and hyper-parameter tuning.

Store-Level Features: We incorporated demographic data from the US Census Bureau to create features of each store's location such as unemployment rate, gender, race, male-ratio etc. These features are important for providing a nuanced understanding of the market dynamics at each store, for example lower income areas might prefer lower priced items, enabling us to make informed predictions about store-specific revenue and inventory costs in the upcoming quarters.

Product-Level Features: We have also focused on the characteristics of each SKU, creating dummy variables such as color and brands. This will enable us to make recommendations on specific SKUs for inventory and promotions. To mitigate the effects of price fluctuations over time, we employed average cost techniques. Smoothing our costs will enable us to measure changes in cost as well as showing improvements in ROI.

Temporal Analysis: Acknowledging the impact of seasonality on consumer purchasing patterns, our data frame is organized by quarters. This temporal aspect allows us to model and predict seasonal trends, thereby anticipating the popularity of certain SKUs during specific seasons, which can offer recommendations for pricing, marketing, and inventory allocation.

Expected Outcomes and Applications: Our model is expected to yield accurate forecasts of quarterly revenue predictions at both the store and SKU levels, accounting for various demographic and product-specific factors. The insights from our analysis will be instrumental in inventory planning, targeted marketing, and overall strategic planning, leading to optimized resource allocation and improved ROI.

4.Model Evaluation

In the machine learning segment of our project, we undertook a randomly sampled data frame as a subset from the master data frame comprising of 320,000 rows and 1,205 variables, followed by a validation process on a subset of 80,000 rows. The target variable, Revenue, underwent a log transformation to arrive at a normal distribution, enhancing the applicability of linear models. Our feature reduction efforts were guided by the goal of curbing overfitting. By systematically eliminating non-contributing numeric features, we refined our model only to include variables with substantial predictive power. We utilized random sampling to partition our data into training and validation sets, ensuring a robust framework for model assessment. Our model suite included Linear Regression, Ridge Regression, Lasso Regression, and XGBoost, each evaluated using R-squared and Mean Squared Error (MSE) as the primary metrics for performance.

Model	Train Set MSE	Train Set R^2	Validation Set MSE	Validation Set R^2
Linear Regression	0.329	0.661	0.329	0.662
Ridge Regression	0.329	0.661	0.329	0.662
Lasso Regression	0.333	0.657	0.332	0.659
XG Boost Regressor	0.267	0.725	0.272	0.72

Our XGBoost Regressor marked a notable improvement, with the training set exhibiting an MSE of 0.267 and an R2 of 0.725, whereas the validation set recorded an MSE of 0.272 and an R2 of 0.72. This enhancement in the metrics indicated a substantial leap in predictive accuracy and model robustness. Hyperparameter tuning was conducted on the XGBoost model to optimize its performance metrics, resulting in the following tuned values: 'colsample_bytree' set to 1.0, 'learning_rate' to 0.3, 'max_depth' to 9, 'n_estimators' to 200, and 'subsample' to 1.0. This fine-tuning aimed to optimize the model's complexity and fit to the data, ensuring the most accurate predictions possible.

5.Model Assumptions for Business Results

Our analysis focuses exclusively on SKUs that recorded sales in both the previous and current quarters. The baseline model operates under the assumption that the quantity of sales for each SKU will mirror that of the previous quarter. In contrast, our predictive model estimates both the revenue and quantity sold for each SKU, using a trained algorithm and various features. We assess overstocked SKUs as identified by both our model and the baseline model. This enables us to determine the financial impact resulting from the overstocked inventory. Our predictions on Revenue bolster a return-on-investment driven by reducing costs through proper inventory management and cost reduction.

6.Business Summary

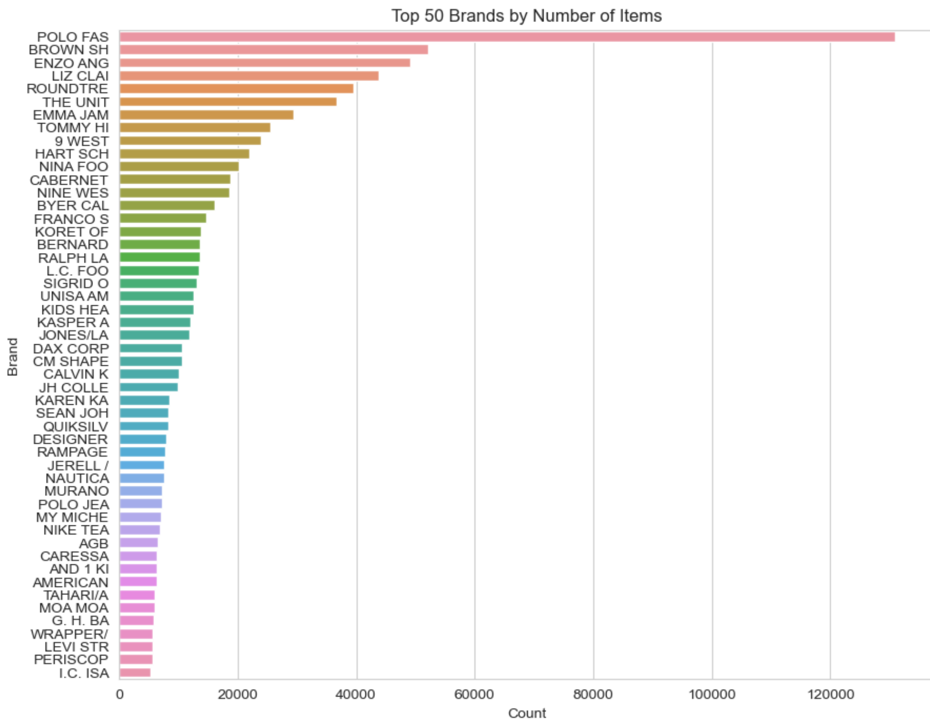
The primary objective was to enhance Return on Investment by reducing inventory costs. In 2005, the model reduced inventory overspending, saving $16.9 million in Q1, $15.7 million in Q2, and $23.5 million in Q3. Overall, the model contributed to a total reduction in inventory costs of $117 million for the year. The model's success is evident in its reduction of inventory costs, achieving a significant ROI of 21.6% for the year 2005, inclusive of data science costs.

7.Implications

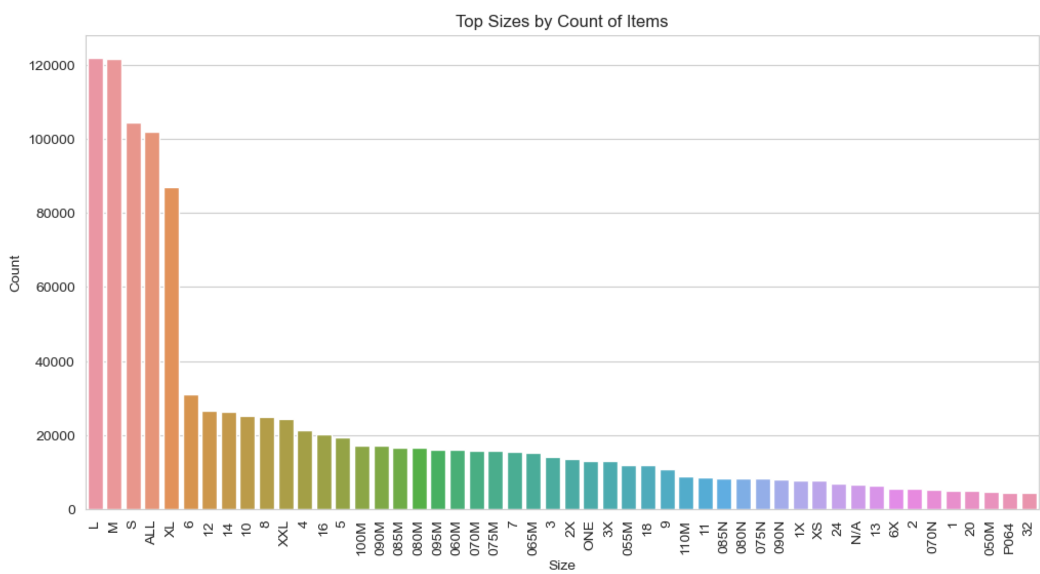
Overstocking can significantly disrupt a business's operations and financial health. It results in increased storage costs due to the need for additional space, which escalates expenses like rent, utilities, and maintenance. Capital, rather than being invested in growth opportunities, becomes tied up in unsold inventory. This excessive stock, particularly in fast-paced industries, may face obsolescence, while environmentally sensitive items risk deterioration and damage. Consequently, cash flow is impacted negatively as funds are locked in non-revenue-generating goods. To mitigate overstocking, Dillards may implement discounts and markdowns, which, while clearing inventory, also compress profit margins. Moreover, overstocking signifies an inefficient allocation of resources such as manpower and equipment, which could be employed more effectively in other areas of the business. Supply chain complications can arise as well, with excess inventory causing logistical disruptions. Furthermore, overstocking can inadvertently lead to a diminished perception of product quality among customers, especially if items are frequently seen on clearance. Lastly, the opportunity cost of overstocking cannot be overlooked; the resources and space consumed by surplus stock could have been allocated to more in-demand products, potentially leading to lost sales and market opportunities.

The variation in ROI especially in Q2, where it fell compared to Q1 and Q3, highlight the need for enhanced revenue forecasting. Improving the model's ability to anticipate and adapt to market trends and consumer behavior could lead to even more significant financial benefits. This suggests a clear direction for future development: focusing on advanced algorithms and data analysis techniques that can provide deeper insights into revenue generation patterns. Our confidence in the model's capacity to deliver substantial ROI benefits is rooted in its proven ability to control and optimize costs. However, we acknowledge that revenue management, often subject to external market variables and consumer trends, presents a complex challenge. By bolstering the model's predictive accuracy in revenue streams, we can not only control costs but also proactively influence revenue, further enhancing the financial performance of Dillard's organization. This dual focus on cost management and revenue optimization positions our model not just as a forecasting tool, but as a comprehensive strategic asset for financial efficiency and profitability.

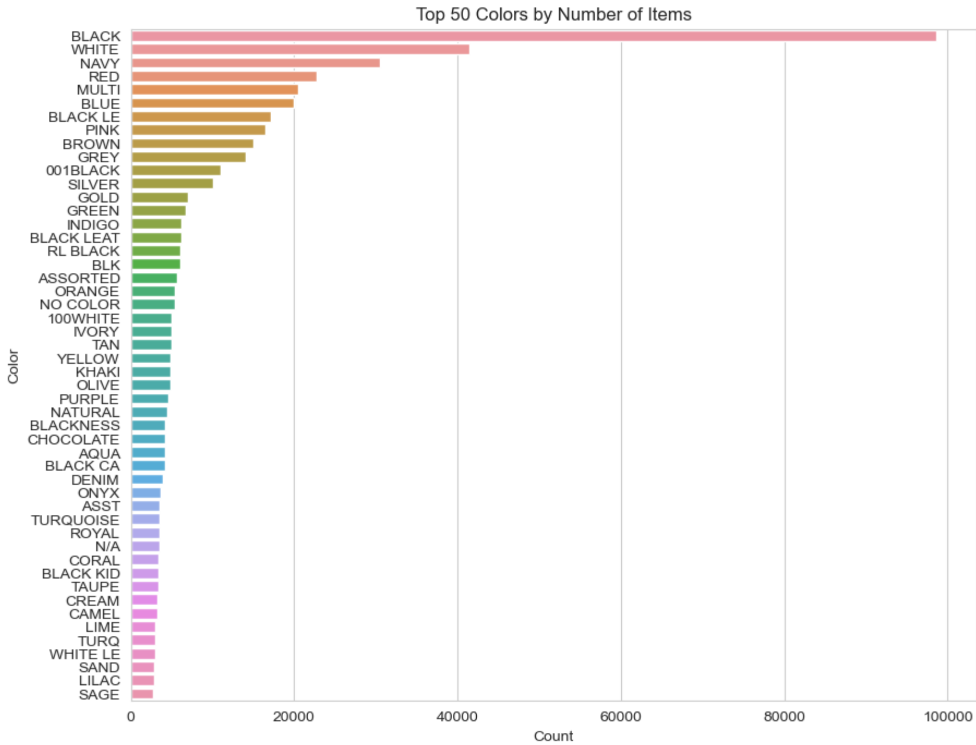
Appendix



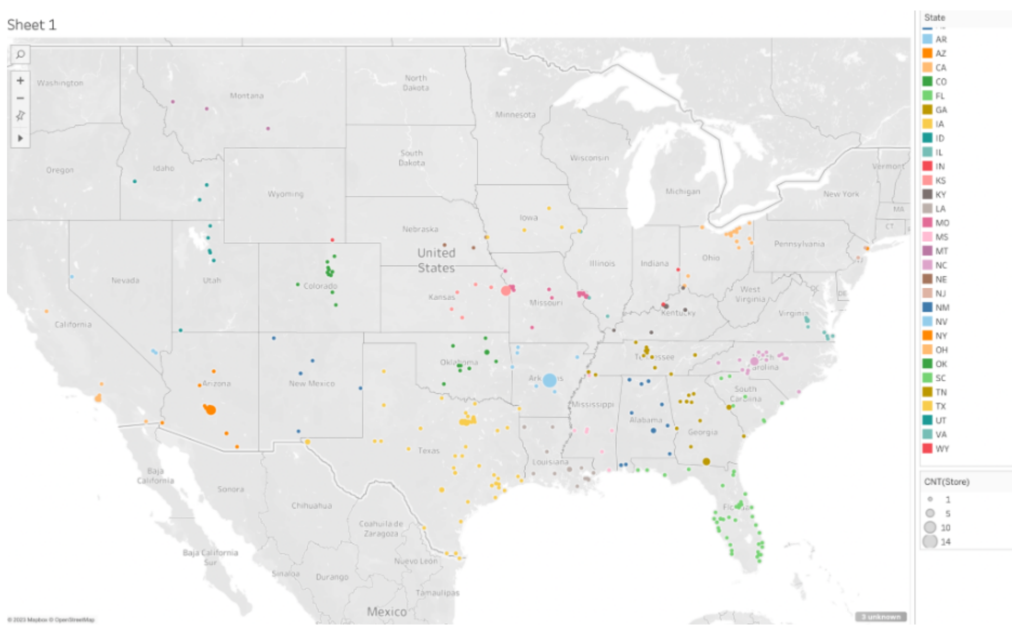
1. Top 50 Brands



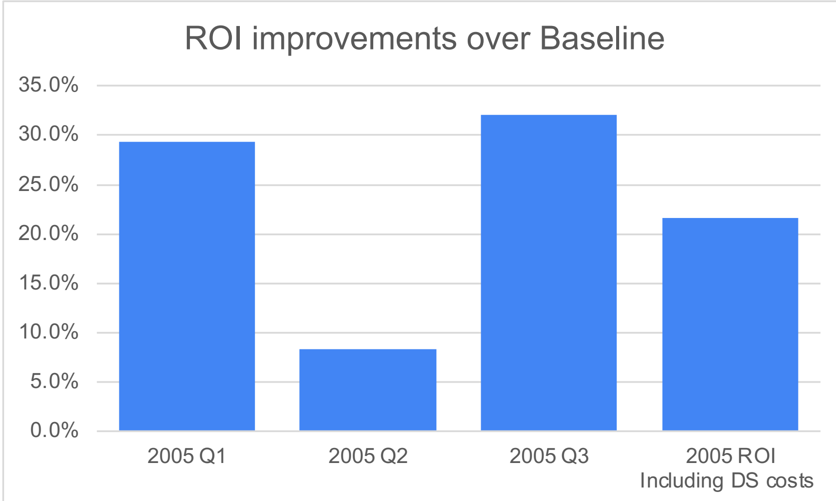
2. Top Sizes



3. Top Colors



4. Heat map of store locations



5. ROI Improvements Q/Q and 2005 total

2005 Q1		
Actual Revenue	$	195,679,558.66
Baseline Inventory	$	74,200,000
Model Inventory	$	16,900,000
Dollars Saved	$	57,300,000.00
ROI Improvement (Over Baseline)		**29.3%**
2005 Q2		
Actual Revenue	$	215,510,342.51
Baseline Inventory	$	33,700,000
Model Inventory	$	15,700,000
Dollars Saved	$	18,000,000.00
ROI Improvement (Over Baseline)		**8.4%**
2005 Q3		
Actual Revenue	$	130,828,746.49
Baseline Inventory	$	65,500,000
Model Inventory	$	23,505,000
Dollars Saved	$	41,995,000.00
ROI Improvement (Over Baseline)		**32.1%**
Total Actual Revenue	$	542,018,647.66
Total Inventory Cost Savings	$	117,295,000.00
Total ROI Improvement 2005		**21.640%**
Data Scientist Cost		
Durations (days)		180
FTE		2
Annual salary	$	200,000.00
Pay	$	200,000.00
Computing hours	$	500.00
Cloud per hour	$	10.00
Cloud Amount	$	5,000.00
Total Cost	**$**	**205,000.00**
Dollars saved in Inventory after labor costs	**$**	**117,090,000.00**
2005 ROI Including DS costs		**21.603%**

6. ROI Analysis