Wholesale Customer Analysis

Johnny Lee

March 31, 2025

Contents

1 Introduction

The Wholesale Customer dataset originates from a wholesale distributor in Portugal and captures the annual purchasing behavior of 440 customers. The dataset consists of both categorical and continuous variables that provide valuable insights into customers' purchasing patterns. The categorical variables include `Channel`, which classifies customers into two groups: `Horeca` for Hotels, Restaurants, and Cafes, and `Retail` for Retailers, such as grocery stores. The second categorical variable, `Region`, categorizes customers based on their location in Portugal, with three possible values: `Lisbon`, `Oporto`, and `Other` for other regions in Portugal that aren't Lisbon and Oporto.

The continuous variables in the dataset represent the annual spending in monetary units across 6 product categories. These include `Fresh`, which tracks spending on perishable food items like fruits, vegetables, and meats; `Frozen`, for spending on frozen food items; `Milk`, which indicates spending on dairy products; `Grocery`, which reflects spending on non-perishable food items; `Detergents & Paper`, for cleaning supplies; and `Delicatessen`, which accounts for spending on delicatessen products such as cheeses and prepared foods.

This analysis aims to answer two key questions: first, whether it is possible to predict a customer's grocery spending based on other purchasing patterns, and second, whether a customer's channel can be determined from their spending behavior. These questions are of particular interest, as understanding the relationships between these variables could lead to more efficient operational strategies.

Answering these questions would offer several advantages for the wholesaler. In terms of grocery spending analysis, the ability to forecast grocery spending accurately would facilitate better warehouse planning by ensuring that the appropriate amount of storage space is allocated. It would also help with financial management, as accurate predictions would allow the wholesaler to set aside the necessary funds for inventory needs. Additionally, delivery scheduling could be optimized to align with expected purchase volumes, reducing delays or overstocking. With better insights into grocery spending, the wholesaler could reduce the risks of stockouts and overstocking, while more reliable spending estimates could also lead to better vendor negotiations.

On the other hand, determining the customer channel based on spending behavior provides its own set of benefits. By automating the process of categorizing customers into channels, the wholesaler can eliminate the need for manual research, saving time and ensuring accurate classification. This understanding would also enable more targeted marketing efforts, where promotions and special deals could be tailored based on the customer's channel. Furthermore, assigning dedicated service representatives to each channel would improve customer service, fostering better communication and faster issue resolution.

To address these questions, an exploratory analysis of the dataset will be conducted, followed by building the models. The exploratory analysis will involve examining the distributions of the variables, assessing correlations between variables, performing hypothesis tests, and visualizing the results.

2 Exploratory Analysis

This section investigates the dataset's structure by analyzing summary statistics and variable distributions, examines relationships between features using correlation analysis, and tests the statistical significance of observed patterns through hypothesis testing. These insights directly inform the design of the machine learning models.

2.1 Summary Statistics and Distributions

Initially, the summary statistics of the continuous spending variables (Table 1) were examined.

Table 1: **Summary Statistics of Annual Spending** (Monetary Units)

Statistic	Fresh	Milk	Grocery	Frozen	Detergents & Paper	Delicatessen
Min	3	55	3	25.0	3.0	3.0
1st Qu	3128	1533	2153	742.2	256.8	408.2
Median	8504	3627	4756	1526.0	816.5	965.5
Mean	12000	5796	7951	3071.9	2881.5	1524.9
3rd Qu	16934	7190	10656	3554.2	3922.0	1820.2
Max	112151	73498	92780	60869.0	40827.0	47943.0

All annual spending categories exhibit right-skewed distributions, as their means are larger than their medians. This skewness suggests the presence of high-spending outliers, indicating that non-parametric methods are needed for the analysis.

Customers spend the most on `Fresh` food items, which dominates across all summary metrics: highest mean (12,000), median (8,504), third quartile (16,934), and maximum value (112,151). In contrast, `Delicatessen` has the lowest mean spending (1,524.9) and third quartile (1,820.2), indicating limited investment in items like cheeses and prepared foods. However, `Detergents & Paper` shows even lower median spending (816.5) and first quartile (256.8), suggesting retailers prioritize essential food items over non-food essentials like cleaning supplies.

The variability in spending, particularly for `Fresh` (IQR = 13,806) and `Grocery` (IQR = 8,503) highlights diverse customer purchasing patterns, with some clients ordering minimal stock (`Fresh` min. = 3) and others making bulk purchases (`Fresh` max. = 112,151).

To further validate these findings, the density plots of annual spending for all variables were examined, stratified by region and channel. Below, the right-skewed distributions of `Fresh`, `Grocery`, and `Milk` are highlighted.

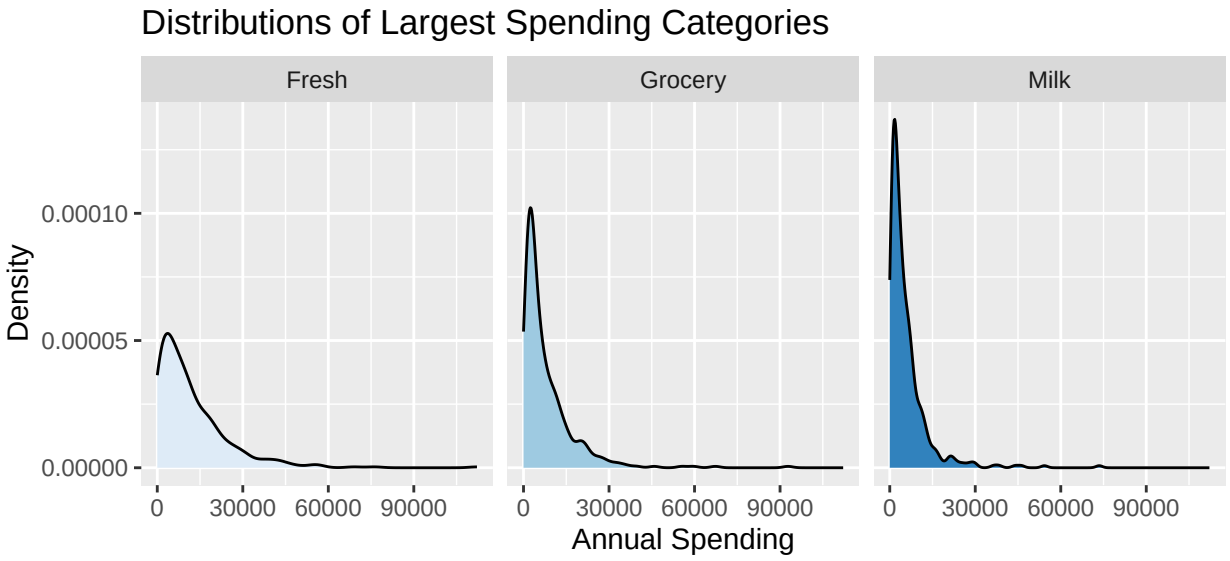

Figure 1: Density distributions of annual spending for Fresh, Grocery, and Milk.

The density plots confirm the right-skewed distributions observed in the summary statistics, with long

tails indicating a small subset of customers making exceptionally large purchases. This reinforces the need for non-parametric methods in subsequent analyses.

2.2 Correlation Analysis

Having examined the distributions of the annual spending variables, the next step was to explore the correlations between them. Given the non-normality of the data, Spearman correlation was used as the statistic (Table 2).

Table 2: **Moderate to Strong Correlations** (Spearman's $\rho \geq |0.3|$)

Feature 1	Feature 2	Spearman's ρ
Fresh	Frozen	0.38
Grocery	Milk	0.77
Detergents & Paper	Milk	0.68
Detergents & Paper	Grocery	0.80
Delicatessen	Milk	0.37
Delicatessen	Grocery	0.30

The correlation analysis identified six significant relationships ($\rho \geq |0.3|$). The strongest correlations occurred between `Grocery` and `Detergents & Paper` ($\rho = 0.8$), `Grocery` and `Milk` ($\rho = 0.77$), and `Detergents & Paper` and `Milk` ($\rho = 0.68$). This suggests customers frequently purchase these categories together, likely as part of routine household grocery shop.

Moderately correlated pairs included `Fresh` and `Frozen` ($\rho = 0.38$), `Delicatessen` and `Milk` ($\rho = 0.37$), and `Delicatessen` and `Grocery` ($\rho = 0.30$). This indicates a subset of customers purchase Milk, Delicatessen, and Grocery items together, while the link between Fresh and Frozen aligns with traditional purchasing of perishables

To visualize these relationships, scatter-plots for all pairs with $\rho \geq |0.3|$ were generated, focusing on the strongest associations: `Grocery` vs. `Detergents & Paper` and `Grocery` vs. `Milk`.

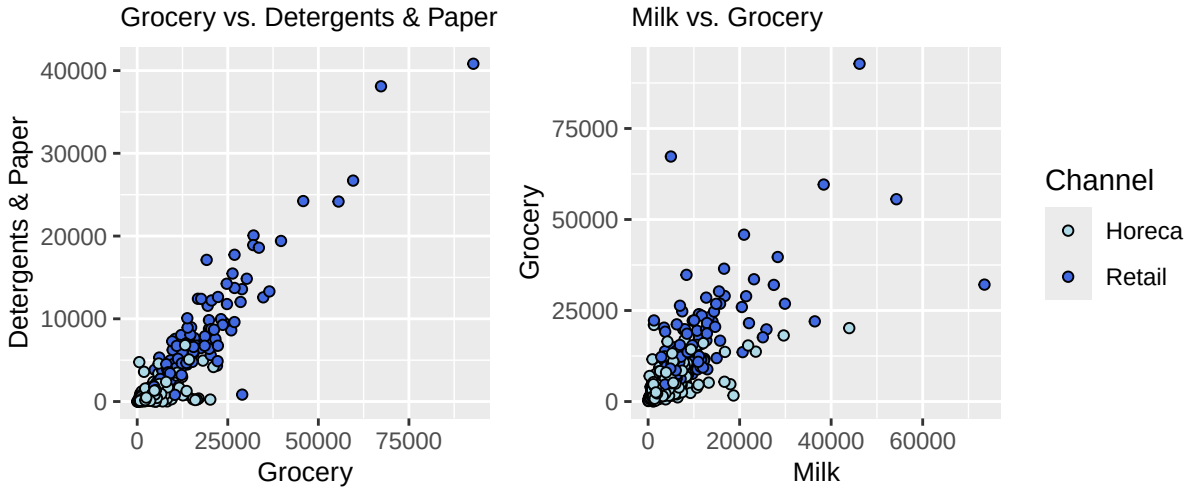

Figure 2: Scatter-plots of Annual Spending: Grocery vs. Detergents & Paper and Grocery vs. Milk

The scatter-plots reveal a strong linear relationship between `Grocery` and `Detergents & Paper`, along with a moderate linear association between `Grocery` and `Milk`. These visualizations reinforce the trends

identified in the correlation analysis, confirming that customers often purchase these product categories together.

2.3 Hypothesis Testing

After examining the correlations, hypothesis testing was conducted. Kruskal-Wallis tests were performed on all six annual spending categories and total annual spending (the sum of all categories) to determine whether median spending differed between regions. These tests revealed no statistically significant differences in median spending across regions at a significance level of $\alpha = 0.05$. Additionally, a Chi-squared test indicated no statistically significant dependence between `Channel` and `Region` (p=value 0.114 for $\alpha = 0.05$).

Next, Mann-Whitney U tests were conducted to compare the median annual spending between the two channels. The results were statistically significant for all six spending categories, as well as for total annual spending ($\alpha = 0.05$), highlighting distinct purchasing patterns between the channels. The p-values for the Mann-Whitney U tests are presented in Table 3 below.

Table 3: ***p*-values of Mann-Whitney U Tests for Annual Spending Categories** ($\alpha = 0.05$).

Feature	p-value
Fresh	<0.001
Milk	<0.001
Grocery	<0.001
Frozen	<0.001
Detergents & Paper	<0.001
Delicatessen	<0.001
Total Stock	<0.001

The table indicates that all p-values are less than 0.001, demonstrating that the differences between the channels are highly significant. This provides strong evidence that the channels differ considerably in all spending categories.

To visually illustrate the differences in median spending between the two channels, box-plots were generated. Below are the two most statistically significant results from the Mann-Whitney U tests are highlighted.

Figure 3: Box-plots of Grocery and Detergents & Paper Annual Spending by Channel

The box-plots clearly demonstrate a substantial difference in medians between the two channels for `Grocery` and `Detergents & Paper` annual spending. Notably, `Retail` exhibits considerably higher spending than channel `Horeca`. These visualizations reinforce the statistical findings from the Mann-Whitney U tests, confirming the distinct purchasing patterns between the channels.

3 Methodology

Given the statistically significant differences in median spending across all categories between customer channels (Mann-Whitney U test, $p < 0.001$ for all variables), classification models were developed to distinguish between the two customer types: `Horeca` and `Retail`. Additionally, the correlation analysis revealed that `Milk` and `Detergents & Paper` were both strongly correlated with `Grocery`, making regression a logical approach for predicting annual `Grocery` spending. To ensure reproducibility, a random seed of 7890682 was set for all models. The modelling approach is outlined below.

3.1 Classification Models

To establish a baseline for comparison, a logistic regression model was built. Logistic regression is a generalized linear model that uses the logit function, transforming the linear combination of predictors into probabilities through an exponential power. This model is valuable for its simplicity and efficiency, and it trains quickly, which is particularly useful when scaling data. Additionally, logistic regression offers high interpretability, as it allows assessment of the contribution of each feature through the magnitude of its coefficients. However, logistic regression comes with some key assumptions, such as the need for a linear relationship between the logit and predictors, no multicollinearity between features, and the absence of outliers. Upon examination of the data, two of these assumptions were found to be violated. There were notable outliers, as revealed by the distributions in Section 2.1, and several features exhibited multicollinearity, as identified in the correlation analysis in Section 2.3. As a result, some observations were perfectly predicted with probabilities of 0 or 1. While these issues could be addressed by removing outliers and highly correlated features, an alternative approach was to explore non-parametric models to avoid these assumptions.

Next, a k-Nearest Neighbors (KNN) model was implemented. KNN is a non-parametric method that makes no assumptions about the underlying distribution of the data. Instead, it classifies an observation by examining the 'k' closest data points in the feature space and assigns the majority class from the nearest neighbors. One of the key advantages of KNN is its simplicity and the lack of model assumptions, making it suitable for the data, given the multicollinearity issues encountered. Since KNN relies on distance calculations, features were scaled using Z-score normalization to ensure that larger-scale features do not disproportionately affect the distance metric. This allows all features to contribute equally to the classification process.

Finally, a random forest model was implemented. A random forest creates an ensemble of decision trees, where each tree is trained on a bootstrapped subset of the data and considers a random selection of features at each decision node. This randomization ensures that each tree is slightly different, promoting diversity in the model and improving its generalization capabilities. In classification, the prediction is made by aggregating the results from all trees in the forest, with the final classification being the majority vote. While individual decision trees are highly interpretable due to their clear decision rules, random forests as an ensemble method, are less interpretable. However, this complexity is offset by their superior performance, as they tend to yield more accurate predictions compared to simpler models. A random forest is more complex than logistic regression and KNN, but it is likely to provide the best performance, as it effectively handles non-linear relationships and reduces the impact of overfitting through its ensemble approach.

To validate the models, a 70/30 train/test split was applied. Additionally, 10-fold cross-validation was used, repeated across all models, to enhance generalizability and mitigate the risk of results being influenced by chance. Given that the dataset was approximately 63% `Horeca` and 37% `Retail`, up-sampling for `Retail`

was implemented to ensure balanced class representation, training the models with a 50/50 distribution between both channels. In 10-fold cross-validation, the dataset is partitioned into 10 roughly equal subsets. One subset is used for testing, while the other nine are used for training. This process is repeated such that each fold is used as a test set exactly once. The performance metrics are averaged across the 10 iterations to assess the model's generalizability and confirm that the results are not a product of random variation. Model performance was evaluated using accuracy, ROC-AUC, PR-AUC, precision, recall, and F1-score, with ROC-AUC being the most critical to ensure strong overall classification performance across all channels.

Finally, the hyper-parameters for the KNN and random forest models were fine-tuned. For the KNN model, various values for the number of neighbors, $k \in \{3, 5, \ldots, 17\}$ were tested. For the random forest model, the following parameters were explored: the number of trees $t \in \{100, 200, \ldots, 500\}$, the number of features to randomly select at each decision node $f \in \{2, 3, 4, 5\}$, and the minimum number of observations required in a node $m \in \{1, 5, 10\}$. This grid search enabled the identification of the best-preforming parameters for each model.

3.2 Regression Models

Similarly to the classification models, a baseline regression model was first constructed for comparison. This baseline model was simply the mean annual grocery spending from the training data.

Next, feature engineering was performed to determine which predictors to include in the regression models. Initially, `Detergents & Paper` was the only feature, as it was the most strongly correlated with `Grocery`. Then, `Milk` was added, as it was the second most correlated feature. After that, different feature combinations were experimented with, which included a model with all seven features and one with only the continuous features. Finally, the relationships of the 5 continuous features with `Grocery` were revisited by generating a pairs plot of the 6 continuous variables (including `Grocery`).

The pairs plot suggested that `Fresh`, `Frozen`, and `Delicatessen` might not have a strictly linear relationship with `Grocery`, so degree 2 and 3 polynomial terms for these features were incorporated. The plot also reaffirmed the presence of multicollinearity among the features, prompting the addition of interaction terms for `Fresh` and `Frozen`, `Detergents & Paper` and `Milk`, as well as `Milk` and `Delicatessen`.

Next, regularized regression was applied to models with many correlated features. Two regularization methods were employed: ridge and lasso regression. Ridge regression shrinks coefficients by adding the L_2 norm of the coefficients to the Ordinary Least Squares loss function, reducing coefficient magnitude without driving them to zero. This approach is ideal for shrinking correlated features without excluding them entirely. In contrast, lasso regression incorporates the L_1 norm of the coefficients into the Ordinary Least Squares loss function, which not only shrinks coefficients but can also reduce them to zero, effectively performing feature selection. These regularization techniques were applied to models with more than two features, including those with interaction and polynomial terms, and the model with continuous features only.

As with the classification models, the regression models were validated using a 70/30 train/test split. In addition, 10-fold cross-validation was applied (as explained in Section 3.1) across all models to enhance generalizability and reduce the risk of chance findings. Model performance was evaluated using RMSE.

Finally, the λ values in the ridge and lasso regression models were fine-tuned to optimize for RMSE. Additionally, the coefficients were analyzed for how they shrank as λ increased, to identify which features were most influential in the models.

4 Results

This section discusses the results of the classification and regression models in predicting customer channel and grocery spending. The performance of the classification models was quantified using ROC-AUC, while the regression models were evaluated based on RMSE. To reiterate, all models underwent 10-fold cross-validation to ensure that the findings were not due to chance.

4.1 Classification

Firstly, the logistic regression model will be discussed, with a particular focus on its key performance metric, the ROC-AUC. The model achieved an impressive test ROC-AUC of 0.936 and a cross-validation ROC-AUC of 0.958, both of which are excellent results, especially for a baseline model. The model was trained on z-score scaled data to allow for a fair comparison of the magnitudes of the feature coefficients. Table 4 below highlights the coefficients of the logistic regression model, which reflect the relative importance of each feature in the scaled data.

Table 4: **Logistic Regression Coefficients** (Standardized Coefficients)

Feature	Coefficient	Std. Error	p-value
(Intercept)	-2.645	0.730	0.000
Region Oporto	2.524	0.977	0.010
Region Other	2.317	0.725	0.001
Fresh	0.359	0.260	0.168
Milk	0.974	0.404	0.016
Grocery	0.447	0.683	0.513
Frozen	-2.568	0.769	0.001
Detergents & Paper	4.713	0.777	0.000
Delicatessen	0.060	0.378	0.873

The results highlight `Detergents & Paper` as the most influential predictor in the logistic regression model, with the largest absolute coefficient ($|\beta_{\texttt{Detergents \& Paper}}| = 4.713$) and near-zero p-value ($p \approx 0$), indicating exceptional statistical significance. Other statistically significant predictors, ranked from most to least significant, include the intercept ($|\beta_0| = -2.645, p \approx 0$), `Frozen` ($|\beta_{\texttt{Frozen}}| = 2.568, p = 0.001$), `Region Oporto` ($|\beta_{\texttt{Region Oporto}}| = 2.524, p = 0.010$), `Region Other` ($|\beta_{\texttt{Region Other}}| = 2.317, p = 0.001$), and `Milk` ($|\beta_{\texttt{Milk}}| = 0.974, p = 0.016$).

Notably, while the earlier Chi-squared test found no direct link between `Channel` and `Region` ($p = 0.114$), the logistic regression model reported `Region` as statistically significant ($p = 0.010$ and $p = 0.001$). These results are contradictory.

This mismatch suggests the logistic regression model might reflect random chance in the data rather than a true connection between `Channel` and `Region`. However, these effects are likely negligible in comparison to the dominant predictor, `Detergents & Paper`, which has the largest coefficient and a near-zero p-value. `Detergents & Paper` is likely the most significant feature because Retailers buy more non-food items like cleaning supplies than `Horeca` customers.

Next, the results of the KNN model will be discussed. Like the logistic regression model, the KNN model was trained on z-score scaled data (as explained in Section 3.1). However, unlike logistic regression, there is no explicit model in KNN. Instead, the majority class among the k nearest neighbors is used to classify new data points. The grid search identified the optimal number of neighbours as $k = 17$, the highest tested value, suggesting that incorporating more neighbours improves generalization by reducing sensitivity to local noise.

The KNN model achieved a test ROC-AUC of 0.944, slightly outperforming logistic regression (0.936), but its cross-validation ROC-AUC (0.948) was marginally lower than logistic regression's (0.958). Given the minimal difference in performance metrics, the two models are comparable in predictive power. However, the lack of additional metrics makes it challenging to definitively declare one model superior.

Finally, the results of the random forest model are discussed. Unlike KNN and logistic regression, the random forest was trained on the unscaled original data, as tree-based models do not rely on distance metrics or coefficient interpretation, rendering scaling unnecessary. The grid search identified the optimal hyper-parameters as $t = 500$ decision trees, $f = 2$ randomly selected features per split, and a minimum node size of $m = 10$ observations. The large number of trees ensures predictions stabilize through aggregation, while limiting features per split reduces tree correlation, promoting generalization. The minimum node size balances simplicity and predictive power, removing overly complex splits to mitigate overfitting.

The random forest model achieved a test ROC-AUC of 0.955, marginally outperforming KNN (0.945) and logistic regression (0.936), and its cross-validation ROC-AUC of 0.970 surpassed both counterparts (KNN: 0.948, logistic regression: 0.958).

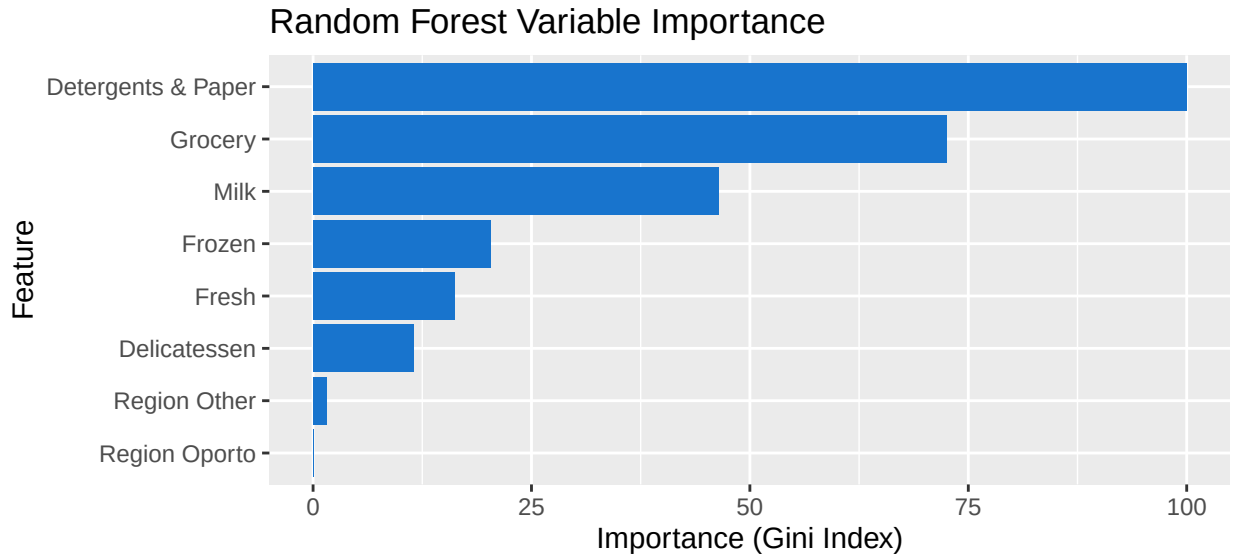

Figure 4: Random Forest Feature Importance (Gini Index)

Figure 4 highlights the importance of each feature according to the Gini Index (G), which quantifies a feature's contribution to node purity across the forest. The top predictors were `Detergents & Paper` ($G = 100.0$), `Grocery` ($G = 72.5$), and `Milk` ($G = 46.4$), reinforcing their role in distinguishing Retailers from `Horeca` customers (as found in Section 2.3). Notably, `Region` had zero importance ($G_{\texttt{Other}} = 1.56$, $G_{\texttt{Oporto}} \approx 0$), aligning with the earlier Chi-squared test ($p = 0.114$), which found no direct dependence between `Channel` and `Region`. However, this doesn't match up with what was seen earlier with the logistic model, where Region did play a role ($p = 0.010$ and $p = 0.001$). This reinforces that the significant `Region` coefficient was likely due to random chance.

While these metrics position the random forest as the top performer, the narrow margins suggest all three models are broadly comparable. Without additional metrics, it remains challenging to declare a single superior model.

Now, the models can be compared. Below, Table 5 summarizes the performance of the classification models. It includes accuracy, recall, precision, F1-score, ROC-AUC, and PR-AUC for all three models (logistic regression, KNN, and the random forest), along with their cross-validation performance metrics, denoted by the model name followed by 'CV'.

Table 5: **Model Performance Comparison**

Metrics	Logistic Test	Logistic CV	KNN Test	KNN CV	Random Forest Test	Random Forest CV
Accuracy	0.909	0.910	0.871	0.899	0.924	0.925
Recall	0.939	0.913	0.866	0.879	0.963	0.926
Precision	0.917	0.957	0.922	0.976	0.919	0.968
F	0.928	0.933	0.893	0.924	0.940	0.945
ROC AUC	0.936	0.958	0.944	0.948	0.955	0.970
PR AUC	0.913	0.930	0.963	0.464	0.971	0.907

The random forest model emerged as the superior classifier across nearly all performance metrics, demonstrating robust predictive power and generalizability. The random forest model outperformed logistic

regression (0.924 vs. 0.909) and KNN (0.924 vs. 0.871) in both test accuracy and cross-validation accuracy (0.925 vs. 0.910 for logistic regression and 0.925 vs. 0.899 for KNN), with logistic regression following closely behind. Notably, the random forest model dominated recall, critical for minimizing false negatives, achieving a test score of 0.963 against logistic regression's 0.939 and KNN's 0.871, while maintaining strong cross-validation recall at 0.926. KNN won the precision category with test and cross-validation precision values of 0.922 and 0.976, respectively, but the random forest was close behind with 0.919 and 0.968 for test and cross-validation, showing that both models performed very similarly. The random forest also excelled in balancing precision and recall, securing the highest F1 scores of 0.940 on the test set and 0.945 in cross-validation. As the primary metric, ROC-AUC highlighted the random forest's superiority with test and cross-validation scores of 0.955 and 0.970, respectively, outperforming logistic regression (0.936 and 0.958) and KNN (0.944 and 0.948). While logistic regression showed stability in PR-AUC with test and cross-validation scores of 0.913 and 0.930, the random forest led on the test set with 0.971, whereas KNN suffered severe overfitting with a cross-validation PR-AUC of 0.464. Combined, these results position the random forest as the optimal choice due to its minimal performance gaps between test and cross-validation, dominance in ROC-AUC for class separation, and balanced performance across metrics.

To further evaluate model performance, the PR and ROC curves are examined. These curves visually represent the previously discussed ROC-AUC and PR-AUC metrics. Below are the PR and ROC curves.

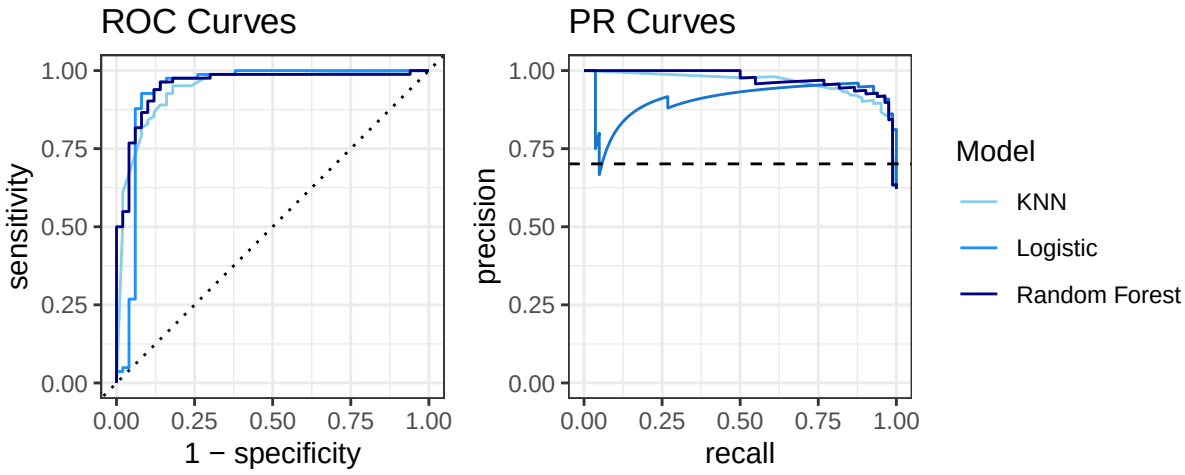

Figure 5: ROC and PR curves of the Classification Models (Logistic Regression, KNN, Random Forest)

The ROC curves for all three models are quite similar, suggesting comparable overall classification performance. However, the random forest model appears to have a slight edge, as its curve maintains a larger area under the curve near top-left corner of the plot. A similar trend is observed in the PR curves, where all three models perform similarly overall. However, logistic regression shows a noticeable dip below the horizontal dashed line (representing random guessing) at lower recall thresholds, indicating performance worse than random guessing in that region. KNN exhibits the most consistent curve with no sharp drops, while the random forest curve follows closely behind. These visualizations provide further insight into the PR-AUC and ROC-AUC values presented in Table 5.

Overall, the random forest is the clear winner, achieving the highest ROC-AUC (0.955 test / 0.970 CV), accuracy (0.924 test / 0.925 CV), recall (0.963 test / 0.926 CV), and F1-scores (0.940 test / 0.945 CV). The combined high cross-validation and test metrics demonstrate the random forest's stability. In contrast, the KNN model exhibits overfitting, with a very low cross-validation PR-AUC of 0.464, which is far below its test performance of 0.963. Though logistic regression offers interpretability, the random forest's performance metrics and resistance to overfitting solidify it as the optimal choice for channel classification.

4.2 Regression

The initial model served to establish a benchmark for evaluating the performance of the other regression models. The baseline model was the mean annual grocery spending of the training set, which resulted in an RMSE of 11,522 on the test data and a cross-validation RMSE of 7,945. The primary objective of the subsequent regression models is to significantly reduce the RMSE.

Given the strong correlation between `Detergents & Paper` and `Grocery` spending ($\rho = 0.80$), a linear model using `Detergents & Paper` as the sole predictor was developed. This model significantly outperformed the baseline, achieving a test RMSE of 3,980 and a cross-validation RMSE of 3,500.

To further refine predictions, `Milk`, which was the second-most correlated feature ($\rho = 0.77$), was added to the model. While this reduced the CV RMSE to 3,019, the test RMSE increased to 4,144, revealing a performance gap (ΔRMSE = 1,125) indicative of overfitting. This instability likely arises because `Milk` and `Detergents & Paper` are highly correlated ($\rho = 0.68$), causing the model to rely too heavily on noise in the training data. Consequently, the simpler model with `Detergents & Paper` only generalized more reliably to unseen data.

A model with all 7 features was then created. This model resulted in the lowest test RMSE yet of 3,944 and a cross-validation RMSE of 3093. Again there is a large difference in RMSE values here (ΔRMSE = 851), which may indicate some overfitting. Adding all features into the model increases multi-collinearity, so this must be addressed. But, before that, it was decided to remove `Channel` and `Region` from the model, which resulted in a test RMSE of 3939 and a cross-validation RMSE of 3062, which are the lowest RMSE values yet.

To explore relationships between features and `Grocery` spending, a pairs plot with Spearman correlations was generated.

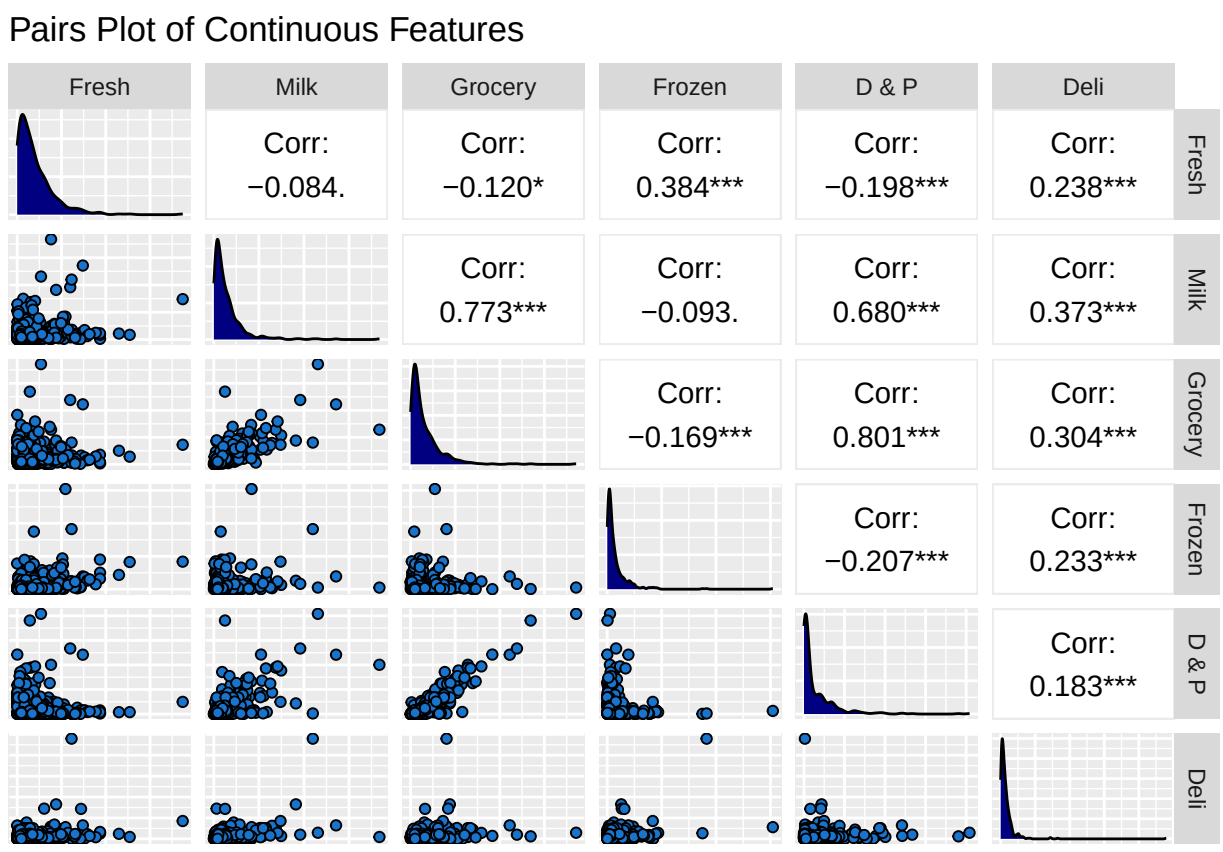

Figure 6: Pairs plot of Continuous Features with Spearman Correlations

The plot reveals potential non-linear patterns between `Grocery` and the features `Frozen`, `Delicatessen`, and `Fresh`, suggesting that polynomial terms could better capture these trends.

The plot also shows a strong correlation between `Milk` and `Detergents & Paper` ($\rho = 0.68$) and moderate correlations between `Frozen` and `Fresh` ($\rho = 0.38$) and `Milk` and `Delicatessen` ($\rho = 0.37$). These strong to moderate correlations suggest that adding interaction terms for these feature pairs may improve model performance by accounting for joint effects.

Following the pairs plot analysis, polynomial terms were introduced to address non-linear relationships: a degree 2 polynomial for `Frozen` and `Delicatessen`, and a degree 3 polynomial for `Fresh`. Interaction terms for `Milk` and `Detergents & Paper`, `Frozen` and `Fresh`, and `Milk` and `Delicatessen` were also included. Surprisingly, these additions significantly degraded performance, resulting in a test RMSE of 4,761 and cross-validation RMSE of 4,488.

Removing the interaction terms improved results, reducing the test RMSE to 4,035 and CV RMSE to 3,540. While this outperformed the polynomial-interaction model, it still lagged behind the simpler continuous-feature model.

Finally, excluding `Channel` and `Region` yielded the most stable model yet, with a test RMSE of 4,002 and CV RMSE of 3,975 (ΔRMSE = 27). Despite this improvement in stability, overall performance remained sub-optimal compared to earlier models, highlighting the difficulty of balancing model complexity with generalization.

To address multicollinearity, reduce overfitting, and enhance model generalization, regularized regression (ridge and lasso) was applied to all models with more than two features. The data was z-score scaled to standardize feature magnitudes, enabling direct comparison of coefficients across predictors.

The first model examined was the most complex one, incorporating all features, interaction terms, and polynomial terms for `Frozen`, `Delicatessen`, and `Fresh`, which replaced their original linear terms. Both ridge and lasso regression were implemented. The ridge model resulted in a test RMSE of 5,353 and a cross-validation RMSE of 3,217, indicating severe overfitting with ΔRMSE = 2,136. Coefficient plots (not shown) suggest that the combination of polynomial and interaction terms was the primary cause.

The lasso model performed slightly better, yielding a test RMSE of 4,444 and a cross-validation RMSE of 3,132, though overfitting remained an issue (ΔRMSE = 1,312). Given that interaction terms have consistently degraded performance across models, they should likely be removed.

To further analyze the impact of regularization, cross-validation MSE vs. log(λ) curves were generated for both models (Figure 7 below).

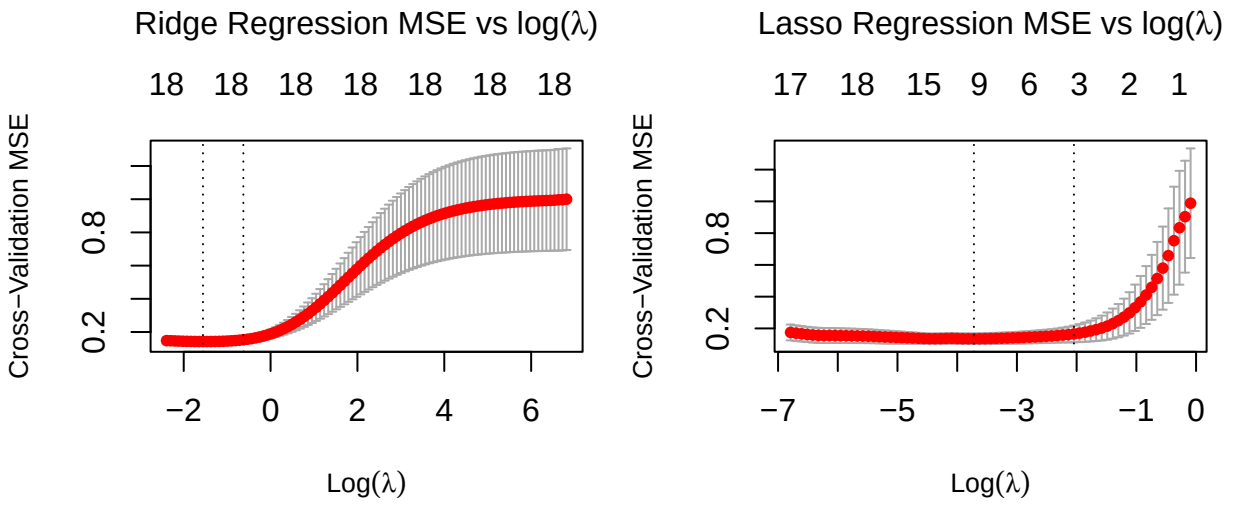

Figure 7: CV MSE vs log(λ) for Ridge/Lasso models with poly (Frozen, Deli, Fresh) and interaction terms.

The first vertical dotted line in the plots indicates the lambda value that resulted in the lowest MSE. Ridge regression benefits slightly from coefficient shrinkage, achieving optimal performance at $\log(\lambda) = -1.557$, while lasso regression shows a greater improvement, with its optimal performance at $\log(\lambda) = -3.72$, as indicated by the first vertical dotted line in its respective plot. This comparison illustrates lasso's ability to leverage the bias-variance trade off by shrinking irrelevant features (reducing variance while increasing bias).

Regularization was also applied to the continuous feature model. The ridge regression model produced a test RMSE of 4,115 and a cross-validation RMSE of 3,292, which represents a significant improvement over the most complex model. The lasso regression model further improved these results, yielding a test RMSE of 3,938 and a cross-validation RMSE of 3,061. This is the best model so far because it achieved the lowest RMSE values. To explore the relationship between MSE and λ, cross-validation MSE vs. $\log(\lambda)$ curves were generated for both models (Figure 8 below).

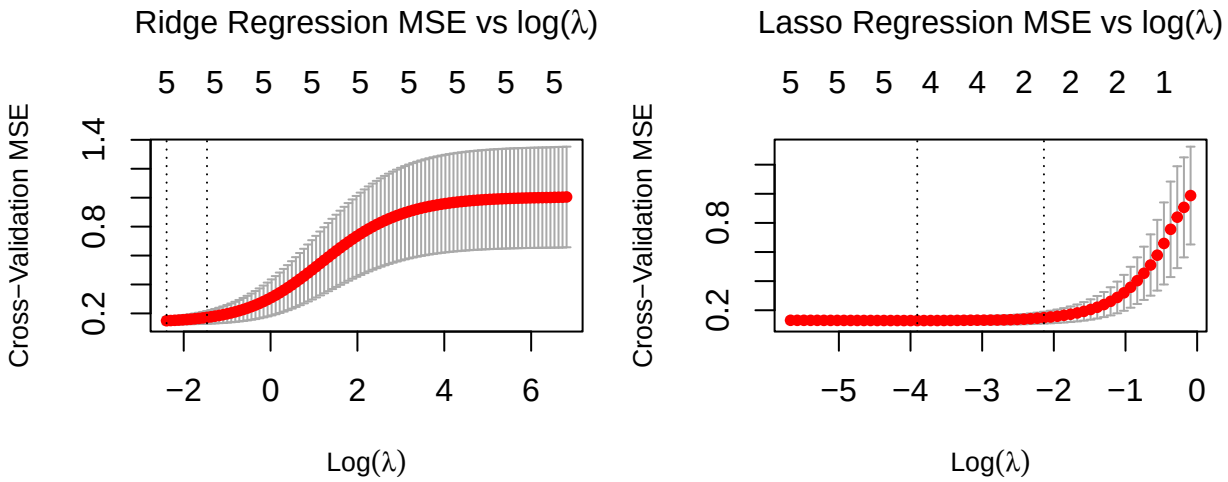

Figure 8: CV MSE vs $\log(\lambda)$ for Continuous Feature Ridge/Lasso Models

The plots indicate that in ridge regression, the best λ value was the lowest one tested, with optimal performance at $\log(\lambda) = -2.39$, as marked by the first vertical dotted line. This clearly demonstrates that any coefficient shrinkage worsened model performance. In contrast, lasso regression benefited from some shrinkage, achieving optimal performance at $\log(\lambda) = -3.91$. This again highlights lasso's capacity to leverage the bias-variance trade-off effectively.

To explore which features affected model performance the most, coefficient path plots were generated (Figure 9 below.)

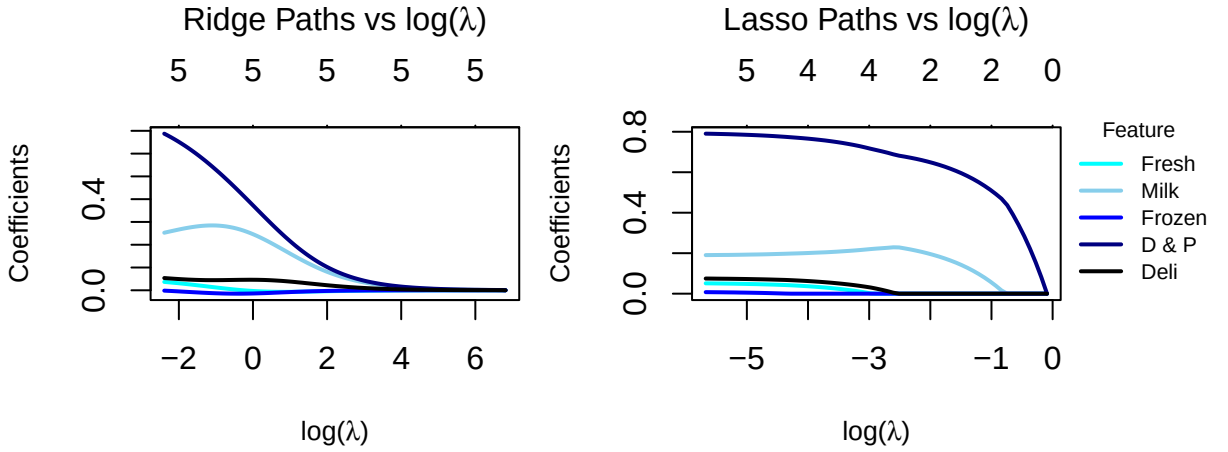

Figure 9: Coefficient Path Plots for Continuous Feature Ridge/Lasso Models

The coefficient path plots indicate that in ridge regression, the most important feature is `Detergents & Paper`, followed closely by `Milk`. As λ increases, all coefficients in the ridge regression model shrink quite rapidly, which may explain why increasing λ resulted in a higher MSE.

Similarly, in the lasso regression model, `Detergents & Paper` remains the most important feature, followed by `Milk`. However, the coefficient for `Detergents & Paper` is significantly larger than that of `Milk`. As λ increases, `Detergents & Paper` barely shrinks, while `Milk` also remains relatively stable but shrinks slightly sooner. This reinforces the idea that the high correlation between `Grocery` and `Detergents & Paper` is the key factor driving model performance.

The last set of regularized regression models was created for the continuous feature model, incorporating polynomial terms for `Frozen`, `Delicatessen`, and `Fresh`, which replaced their original linear terms. The ridge regression model resulted in a test RMSE of 4,165 and a cross-validation RMSE of 3,330. The lasso regression model achieved a test RMSE of 3,956 and a cross-validation RMSE of 3,100. Both models performed worse than the previous continuous feature only model.

Finally, after developing all the regression models, their performance can be compared. Table 6 below summarizes the test and cross-validation (CV) RMSE values for each model.

Table 6: **Model Performance Comparison**

Model	Test RMSE	CV RMSE
1. Base (Mean)	11,522	7,945
2. Detergents & Paper Only	3,980	3,500
3. Detergents & Paper + Milk	4,144	3,019
4. All Features	3,944	3,093
5. Continuous Features Only	3,939	3,062
6. Polynomial + Interactions	4,761	4,488
7. Polynomial (No Interactions)	4,035	3,540
8. Continuous + Polynomial Terms	4,002	3,975
9. Max Feature Ridge	5,353	3,217
10. Max Feature Lasso	4,444	3,132
11. Continuous Ridge	4,115	3,292
12. Continuous Lasso	3,938	3,061
13. Continuous Polynomial Ridge	4,165	3,330
14. Continuous Polynomial Lasso	3,956	3,100

The best-performing model was Model 12, the continuous feature lasso regression model, with a test RMSE of 3,938 and a cross-validation RMSE of 3,061. The most stable model was Model 8, the continuous feature model with polynomial terms replacing the linear terms for `Detergents & Paper`, `Fresh`, and `Frozen`, showing a ΔRMSE of 27. The worst-performing model was Model 1, the base mean model, with a test RMSE of 11,522 and a cross-validation RMSE of 7,945. The most overfit model (excluding the base) was Model 9, the ridge model with substituted polynomial and interaction terms, exhibiting a ΔRMSE of 2,162.

To interpret Model 12's predictions, Table 7 summarizes the z-score scaled model coefficients at the optimal λ value ($\lambda = 0.0201$), selected via cross-validation.

Table 7: **Lasso Regression Z-Score Scaled Coefficients for Model 12** ($\lambda = 0.0201$)

Feature	Coefficient
(Intercept)	0.000
Fresh	0.035
Milk	0.202
Detergents & Paper	0.763
Delicatessen	0.061

Table 7 confirms that `Detergents & Paper` has the strongest positive association with grocery spending ($\beta = 0.763$), followed by `Milk` ($\beta = 0.202$). Lasso regularization shrunk the coefficient for `Frozen` to 0, excluding it from the model, while assigning negligible weights to `Fresh` ($\beta = 0.035$) and `Delicatessen` ($\beta = 0.06$). This aligns with the earlier correlation analysis (Section 2.2), where `Detergents & Paper` ($\rho = 0.80$) and `Milk` ($\rho = 0.77$) showed the strongest relationships with `Grocery`, validating the model's focus on the most predictive features.

Given these results, Model 12 is the optimal choice for predicting annual spending on groceries. This model delivers the best RMSE performance, full interpretability of the retained coefficients, a future-proof design that automatically suppresses irrelevant features if new ones are added, and built-in multicollinearity management, which is crucial for handling correlated features.

5 Conclusion

This analysis addressed two questions for the wholesale distributor: (1) predicting annual grocery spending from other purchasing features, and (2) classifying customer channels (`Retail` vs. `Horeca`) based on spending behavior. Both objectives were answered through exploratory analysis and machine learning modelling.

The random forest model outperformed logistic regression and KNN, the highest ROC-AUC (0.955 test / 0.970 CV), accuracy (0.924 test / 0.925 CV), recall (0.963 test / 0.926 CV), and F1-scores (0.940 test / 0.945 CV). Its ability to handle non-linear relationships and mitigate overfitting through ensemble learning made it the superior choice. Notably, `Detergents & Paper` spending was the most influential predictor, highlighting that `Retail` customers purchase significantly more non-food essentials like cleaning supplies compared to `Horeca` customers. The results enable automated channel classification, which can facilitate targeted marketing and customer issue resolution through dedicated service representatives for each channel.

In predicting grocery spending, lasso regression on continuous features delivered the best performance with the lowest RMSE (3,938 test / 3,061 cross-validation), outperforming all other regression models. The strong correlation between `Grocery` and `Detergents & Paper` ($\rho = 0.80$) drove model performance, with lasso regularization shrinking the coefficient for `Frozen` to zero and assigning negligible weights to `Fresh` ($\beta = 0.035$) and `Delicatessen` ($\beta = 0.06$). This model provides useful forecasts for inventory planning, reducing the risk of stockouts and overstocking.

The skewed distributions and outliers necessitated non-parametric techniques, while regularization addressed overfitting, multicollinearity, and enabled feature selection in regression. Cross-validation proved critical, as interaction terms degraded performance in complex models. Interestingly, while Kruskal-Wallis tests found no regional differences in spending, logistic regression flagged `Region` as statistically significant ($p = 0.010$ and $p = 0.001$), suggesting subtle regional trends warrant further exploration.

Improvements could focus on addressing outliers to reduce overfitting and analyzing grocery spending by channel (`Retail` vs. `Horeca`), as their purchasing patterns differ significantly, as revealed in Section 2.3. Though the Chi-squared test in Section 2.3 found no dependence between `Channel` and `Region`, the logistic regression model contradicted this result, suggesting that further investigation into the relationship between `Channel` and `Region` is warranted.

By using the random forest and lasso regression models, the wholesaler can enhance inventory planning, automate customer segmentation, and make data-driven decisions to optimize operations.