Maven Roasters Sales Analysis

Pop Lucas

[1]Technical University of Cluj-Napoca
poplucas18@gmail.com

Abstract. This document presents an analysis of sales data for Maven Roasters, a coffee shop with three locations in New York. The project utilizes machine learning and statistical techniques to explore sales patterns and consumer preferences. Key objectives include predicting the best-selling product per location, analyzing prices using the Gini index, creating predictive models, and comparing the performance of different algorithms. The dataset contains detailed transaction information, including dates, locations, product categories, and prices. The methodology involves data exploration, preprocessing (data cleaning, encoding categorical variables), and the application of regression (Linear, Random Forest) and classification (Logistic Regression) models. The results indicate that the Random Forest model performed best in predicting sales quantities (R^2=0.97), followed by Logistic Regression for classifying product popularity (accuracy 0.91). The project demonstrates the transformation of raw data into useful information for the coffee shop's management decisions.

Keywords: Sales Analysis · Machine Learning · Predictive Modeling · Maven Roasters · Data Preprocessing · Gini Index · Customer Preferences.

1 Introduction

In this data analysis project, We wish to present how We explored the sales patterns and consumer preferences of Maven Roasters, a coffee shop with three locations in New York. Utilizing machine learning and statistical techniques, in the present documentation, We wish to detail the complete process from initial data exploration to the development of predictive models and the interpretation of the results obtained.

The dataset analyzed contains detailed information about transactions, including dates and timestamps, specific locations, product categories, and product prices. Thus, with the help of this precise and vast information, We had the opportunity to investigate various aspects of the coffee shop's operations, for example: sales trends by location, products, and time periods; identifying factors influencing product prices; developing predictive models for price estimation; dividing products into price categories; analyzing price distribution and balance using the Gini index; and discovering specific consumption preferences for each location.

This documentation details the steps taken to develop this project as part of the Intelligent Systems course. We chose this project because We wished to experiment with applying theoretical knowledge of statistics and machine learning to a real-world dataset. As students somewhat interested in the intersection of technology and business, We considered that this project would offer Us the opportunity to analyze a real business.

1.1 Purpose of the Project

The main purpose of this project is to predict which will be the best-selling product for each location using statistical methods and machine learning algorithms. However, there are also other purposes, for example: price analysis using the Gini index, creation of prediction models, and comparison of algorithms.

2 Data Source and Project Context

This project is based on a dataset containing information about sales made by Maven Roasters, which operates in three locations: Astoria, Hell's Kitchen, and Lower Manhattan. The dataset contains detailed transaction information, including date and time, location, product categories, product types, their details, unit prices, and quantities sold. The data is organized in a CSV file (coffe-shop-sales-revenue.csv), where each line represents an individual transaction.

2.1 Utility of this Dataset

This dataset is ideal for exploring sales trends over time, identifying peak customer traffic days, analyzing the performance of each product, and, of course, identifying factors that contribute to fluctuations in sales volume. Therefore, this dataset provides quite important and vast data that will help Us train the models used.

2.2 Requirements and Objectives

The project has the following main requirements and objectives: **Understanding the data**, which involves examining the distribution of sales across all three locations, identifying relationships between different features, analyzing sales inequality using the Gini index and Lorenz curve, and calculating the informational value (entropy) for each feature; **Modeling and predicting sales**, which involves developing models for predicting sales quantities to identify the best-selling product per location, classifying products into popular or less popular categories, and optimizing model hyperparameters to improve performance and accuracy; **Comparing models**, where the three models will be compared after optimization, and the best one will be chosen; **Explaining results on different instances**, where We will see how well the chosen model predicts on a dataset.

2.3 Desired Outcomes

The desired outcomes from this project are: how sales vary by location, which product categories are most popular, what factors most influence sales quantities, how sales vary by location, a regression model for predicting sales quantities, a classification model for identifying popular products, and an evaluation of the importance of different features in predictions.

In conclusion, the project aims to transform raw transaction data from the CSV file into data that can support sales management decisions for the coffee shop.

3 Related Work

1. **Random Forest**
According to the article [1] published in 2001 by Leo Breiman, it presents an innovative machine learning method based on decision trees. Random Forest represents an ensemble of trees built on subsets of data and features, where each tree is generated using a random selection of predictors at each node, which reduces the correlation between trees and, implicitly, the model's variance. By combining the results of multiple trees, either by majority vote in classification or by averaging in regression, Random Forest is a robust method, resistant to overfitting, capable of handling datasets with a large number of variables, and providing an estimate of their importance. Also, this technique is effective against noise in data and missing values. The studies and experiments presented demonstrate the superior performance of Random Forest compared to other methods existing at the time, making it a reference technique in the field of machine learning. In the presented project, the Random Forest model has the best accuracy, this being 0.97, a fairly high accuracy compared to other tested models.

2. **Linear Regression**
According to [2], the book "Applied Regression Analysis" by Norman R. Draper and Harry Smith is an important work that explains how linear regression works and how it can be used to understand relationships between variables. The authors present basic concepts and show how to build and interpret regression models, offering practical real-life examples. It is explained how to check if your model is good and how to improve results, such as by choosing the right variables or checking for potential errors. We used linear regression in Our analysis and We obtained an accuracy of approximately 0.51, which shows that the model could partially explain the relationships in the data.

3. **Logistic Regression**
According to [3], the book "Applied Logistic Regression" by David W. Hosmer Jr., Stanley Lemeshow, and Rodney X. Sturdivant is one of the most

important works in applied statistics, focusing on logistic regression, an essential method for modeling and predicting binary outcomes, such as "yes" or "no," "success" or "failure." The authors explain in detail the fundamentals of the method, how logistic models are built, and how coefficients are interpreted to understand the impact of explanatory variables on the probability of an event. The book includes numerous practical examples, diagnostic techniques, and strategies for evaluating and validating models, making it accessible and useful for both students and practitioners. Methods for dealing with common problems encountered in real data analysis, such as collinearity or imbalanced data, are also presented. We used logistic regression in Our analysis and We obtained an accuracy of 0.91, which indicates very good performance of the model in correctly predicting outcomes. This book remains a valuable resource for anyone wishing to understand and effectively apply logistic regression techniques in various fields.

4. **Gini Index-Lorenz Curve**
According to the article [4] "A simple method for estimating the Lorenz curve," written by Thitithep Sitthiyot and Kanyarat Holasut and published in Humanities and Social Sciences Communications in 2021, it proposes a simplified and accessible method for estimating the Lorenz curve – an essential tool in analyzing economic inequality. The authors aim to overcome the technical complexity of traditional methods and offer a more intuitive solution that can be easily applied by researchers, students, and specialists who do not have advanced training in mathematics or econometrics.
The Lorenz curve is a graphical representation of the cumulative distribution of income or other resources in a population and is the basis for calculating the Gini index – a numerical indicator of inequality. In the article, the authors present a simple mathematical model, based on logical equations and proportional relationships, which allows the construction of the Lorenz curve starting from aggregated data. Their method reduces data requirements and computational complexity without compromising the accuracy of the estimates.
In one of the graphs displayed in Our project, the application of this method is presented, where the obtained Lorenz curve reflects the cumulative distribution of values for the transaction_id variable, relative to the proportion of the population. The dotted red line represents the line of perfect equality, and the area between it and the real curve is used to calculate the Gini Index, in this case, a value of 0.3334, indicating a moderate level of inequality. Thus, with the help of the method proposed by Sitthiyot and Holasut, We managed to obtain a clear visualization of the data distribution and measure the degree of concentration using Gini.

5. **Entropy**
According to [5], the article "Induction of Decision Trees" by J. Ross Quinlan, published in 1986, is a reference work that underlies the development of modern machine learning algorithms, especially those based on decision

trees. The author proposes a method for constructing decision trees that uses fundamental concepts from information theory, such as entropy and information gain. Entropy is defined as a measure of uncertainty and disorder in a dataset and is used to evaluate how homogeneous the classes of an attribute are. If an attribute divides the data into subsets containing instances of a single class, the entropy is low, and that attribute is considered informative. Information gain measures how much entropy decreases if a certain attribute is used to split the data.
In Our analysis, We applied this fundamental idea by calculating the entropy for columns in a dataset. The results are displayed in a graph, which clearly shows the variations in entropy between attributes.

6. **Data Cleaning**
According to [6], the article "Data Cleaning: Problems and Current Approaches" written by Erhard Rahm and Hong Hai Do, published in 2000 in IEEE Data Engineering Bulletin, offers a clear and structured overview of the challenges and existing solutions in the data cleaning process. The authors highlight that real data is often incomplete, redundant, incorrect, and these problems can significantly affect the quality of results in any data analysis.
The paper emphasizes that the data cleaning process is not just a preliminary stage, whether it involves traditional databases, data integration systems, or data warehouse applications.
Another important point from the article is the emphasis on the difficulty of automating data cleaning, as it largely depends on the specific context of the domain and the quality of data sources.
Within Our analysis, We applied this data cleaning principle to eliminate NaN and NULL values from columns as well as duplicate rows.

7. **Correlation Matrix**
According to [7], the article "Corrgrams: Exploratory Displays for Correlation Matrices" written by Michael Friendly, introduces and develops the concept of "corrgram," a graphical representation designed to help explore and quickly understand the structure of a correlation matrix. The author starts from the idea that, although correlation matrices provide essential information about the relationships between variables.
Friendly proposes various forms of corrgrams that combine color coding, geometric symbols, and variable ordering to highlight important groupings, associations, or patterns among variables.
Within Our analysis, We applied a similar method of visualizing correlations in the form of a heatmap, where the correlation levels between variables are reflected. Thus, the graphical representation of the correlation matrix is an essential practice in exploratory statistical analysis.

8. **Train/Test split**
According to [8], the article "A Critical Look at the Current Train/Test Split

in Machine Learning" written by Jimin Tan, Jjianan Yang, Sai Wu, Gang Chen, and Jake Zhao, analyzes in depth the common practices regarding splitting data into training and test sets in machine learning. The authors emphasize that choosing a fixed proportion, such as 80/20 or 70/30, has become a standard in many studies and projects, but without rigorous justification adapted to the context of each dataset or model.

The paper draws attention to the risks generated by automatically applying such a split, especially when the data is imbalanced or presents non-homogeneous distributions. It is also presented that a fixed split can introduce significant variations in measured performance, thus influencing conclusions about model efficiency. Instead of this static approach, more robust methods such as cross-validation are encouraged, which offer a more stable and correct estimate of the model's real performance.

Within Our analysis, We applied both the traditional data split in an 80% training and 20% test proportion, a very common and easy-to-implement method, as well as cross-validation for evaluating and optimizing models.

9. **Data Preprocessing**

According to [9], the article "Big Data Preprocessing" written by Julian Luengo, Diego Garcia-Gil, Sergi Ramirez-Gallego, Salvador Garcia, and Francisco Herrera, analyzes in depth all essential stages of data preprocessing, with a special emphasis on the challenges encountered when working with very large datasets. The authors emphasize that, in the context of massive volumes of data generated daily, preprocessing becomes a critical and indispensable stage, as raw data is often incomplete and redundant.

The work is structured around the most important components of preprocessing: data cleaning, data transformation, dimensionality reduction, and feature selection. In the cleaning stage, detection and treatment of missing values, elimination of duplicates, and correction of recording errors are discussed. For data transformation, the authors emphasize standardization, normalization, and encoding of categorical variables to allow correct interpretation by algorithms.

Also, an important part of the work is dedicated to reducing data complexity, through techniques for selecting and extracting relevant features. This is necessary to avoid overfitting and to reduce model training time.

In Our analysis, We applied several principles. We cleaned the input data by eliminating missing values and also duplicate rows, which significantly contributed to improving the quality of the models used later, thus demonstrating the practical applicability of the methodologies proposed by the authors of this work.

10. **Histogram**

According to [10], the article "Histograms: A Useful Data Analysis Visualization" written by Regina L. Nuzzo, emphasizes the importance of the histogram as an essential tool in exploratory data analysis. The author highlights that, although it is a simple and often underestimated tool, the histogram offers a clear and intuitive perspective on the distribution of a numerical variable. By dividing values into a series of intervals, the histogram shows the frequency with which these values appear, allowing the user to quickly identify characteristics such as symmetry or asymmetry, the existence of modes (peaks), data spread, and possible extreme values (outliers). In Our analysis, We utilized histograms to examine the distribution of values in numerical columns such as unit_price and transaction_qty. Through these graphical representations, We could observe if the data were normally distributed.

4 Implementation of Theoretical Aspects in the Project

In this chapter, We will detail how theoretical concepts of data analysis, statistics, and machine learning were practically applied in the analysis project.

4.1 Libraries Used and Programming Language

The programming language used for this analysis is Python, and execution is done in a runtime environment called Jupyter.

The main libraries used in Our analysis are **pandas** for manipulating and analyzing tabular data (DataFrames), **numpy** for fast and efficient numerical calculations, **matplotlib.pyplot** for explanatory graphs, **seaborn** for advanced statistical visualizations built on matplotlib, **math** (standard Python module) for basic mathematical functions, **scikit-learn** (sklearn) for machine learning (modeling, preprocessing, evaluation, optimization), **scipy.signal** for filtering and smoothing, **warnings** for managing warnings, and **time** for measuring execution time.

4.2 Database Selection and Preprocessing

We began the analysis by loading the coffee shop sales dataset, eliminating missing and duplicate values to ensure data quality. We transformed the transaction date column into a month format to analyze sales seasonality.

Text (categorical) columns, such as location, product category, type, and product details, were numerically encoded using LabelEncoder to be used later in machine learning models. We aggregated the data at the product, location, and month level, calculating the total quantities sold and the average unit price for each combination.

Thus, We obtained a clean, structured dataset prepared for statistical analysis and predictive modeling, with variables relevant to the project's purpose.

Dataset Description The analyzed dataset contains 149,000 rows representing records from a coffee shop named Maven Roasters, which has 3 locations in New York. Each row in the dataset represents an individual transaction and includes the following information: **transaction_date** representing the date the transaction occurred, **store_location** representing the name of the location where the transaction was made, **store_id** representing the location ID, **product_category** representing the category of the product sold (Coffee, Tea, Bakery), **product_type** representing the product subcategory (Espresso, Latte, Herbal), **product_detail** providing additional data about the product (decaf, organic, etc.), **transaction_qty** representing the number of units sold, **unit_price** representing the price per unit of the product sold, **product_id** representing the product ID, and **transaction_time** the time the transaction was made.

Data Cleaning In the first phase, for better analysis quality and accuracy, We began with a data cleaning stage. We checked for missing values (NULL or NaN) in the dataset to avoid errors in subsequent steps. We also eliminated all duplicate rows so that each transaction is unique and does not incorrectly influence statistics or models. After these operations, We checked again for duplicates or consistency problems.

Through these steps, We obtained a clean dataset, without missing values or duplicates, ready for statistical analysis and modeling.

Data Preprocessing After initial cleaning, We proceeded to data preprocessing to prepare them for analysis and modeling. We extracted the month from the transaction date to analyze sales seasonality. We encoded categorical columns such as location, product category, type, and product details using LabelEncoder so they could be used by machine learning models.

4.3 Data Exploration and Analysis

After cleaning and preprocessing the data, We moved to the exploration and analysis stage to better understand the dataset. We analyzed the distribution of numerical values using histograms, both for the entire dataset and for each location individually. We calculated and visualized the correlation matrix to identify relationships between variables. We also used the Lorenz curve and Gini index to evaluate the inequality of sales distribution across different segments. For categorical and numerical variables, We calculated entropy to measure information diversity.

Through these analyses, We obtained a clear picture of sales distribution, seasonality, and relationships between variables.

4.4 Machine Learning Methodologies Used

Linear Regression We used Linear Regression to predict the exact quantity of products sold for each combination of location, category, type, detail, and month. The model attempted to find a linear relationship between these features and sales volume.

The model provided a basic, fast, and easy-to-interpret prediction. However, performance was limited because the relationships between sales variables are not strictly linear. The R^2 score was quite low, at 0.51.

Logistic Regression We used Logistic Regression to classify products as "popular" or "unpopular," based on their characteristics and context (location, month, type, etc.). Practically, the model predicted the probability that a product would exceed the popularity threshold.

The model had good accuracy, R^2 of 0.91, for this classification task, being useful for identifying products with high sales potential, but not for more complex classifications.

Random Forest Random Forest was the main model for predicting sales quantity and for classifying popular products. We used this model because it can capture complex relationships and interactions between variables, for example, the effect of season, the combination of product type and location.

As for results, it had the best outcomes in Our analysis for both regression and classification, with R^2 being 0.97.

4.5 Implementation of Predictive Models

Model Configuration and Parameters Used **Linear Regression**
The model uses a pipeline that includes data standardization followed by the linear regression itself. We configured the model with **fit_intercept=True**, which allows the model to detect the baseline sales level regardless of feature values. The parameter **positive=False** gives the model flexibility to identify factors that both increase and decrease sales. The performance of this model was moderate, with a score of 0.51, meaning it explains approximately 50% of the variation in sales quantities.

Random Forest
To capture more complex and non-linear relationships, We implemented a Random Forest model. We utilized 200 decision trees (**n_estimators=200**), a sufficient number to ensure prediction stability without excessively increasing computation time. The maximum depth of the trees is 20 (**max_depth=20**). To prevent overfitting, We imposed the condition that each node must contain at least 5 examples. We selected features for splitting using the "sqrt" strategy.

Logistic Regression
To classify products into popular and less popular categories, We implemented a logistic regression model. We configured the model with **class_weight="balanced"** to compensate for differences between class frequencies. The maximum number of iterations was set to 5000 (**max_iter=5000**), ensuring the algorithm has enough time to find the optimal solution.

Model Validation and Training All models were evaluated using a rigorous 5-fold cross-validation approach. This validation method provides a more robust estimate of the models' actual performance on new data.

For regression models, We utilized Mean Squared Error (MSE) and the coefficient of determination (R^2) as evaluation metrics. For classification, We analyzed accuracy, precision, recall, and F1-score.

Training the models highlighted the importance of features related to location and product type, confirming the initial intuition that preferences differ significantly between locations and that certain product categories have clear sales seasons.

4.6 Testing and Validation

1. In the first phase, We separated the data into training (80%) and test (20%) sets using stratification by location to maintain the correct data distribution for each store.
2. For a more solid evaluation, We applied KFold cross-validation with 5 partitions, which gives Us a more stable estimate of performance on new data. The results of this cross-validation (**average R^2=0.85 with standard deviation** $^{+}_{-}0.024$) indicate a stable model with small variations.
3. We utilized multiple metrics to evaluate regression models, these being **Mean Squared Error (MSE)**, which measures the average squared error of predictions, and the **Coefficient of Determination (R^2)**. And for Classification, **Accuracy** and the **Classification Report**.
4. We used performance analysis of models per each location; this type of analysis allowed Us to identify variations in prediction accuracy in different locations.
5. We analyzed feature importance to validate their relevance in the model, thus confirming that location, product type, and seasonality are determining factors.
   Therefore, the testing and validation process demonstrates that the Random Forest model offers the best predictions with an R^2 above 0.80 in cross-validation.

5 Results

Following analyses of the model's predictions for several specific instances from the test set, We can observe how different features influence the estimated sales quantity.

Instance 0 - Astoria, Coffee, Espresso **Main characteristics:** Astoria location, Coffee product type, December month. **Contributions to prediction:** Location contributes the most, followed by product type and month. Unit price has a smaller but significant contribution. **Explanation:** Espresso coffee has high sales in winter due to seasonality.

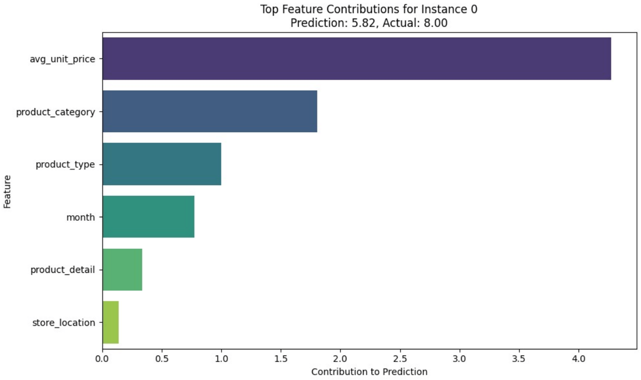

Fig. 1. Instance 0: Astoria, Coffee, Espresso Sales Prediction Factors.

Instance 1 - Hell's Kitchen, Tea, Herbal **Main characteristics:** Hell's Kitchen location, Tea product type, July month. **Contributions to prediction:** Location and product type dominate. **Explanation:** In business areas, cappuccino has stable sales throughout the year. (Note: The explanation seems to refer to cappuccino, while the instance is Herbal Tea. This might be a mismatch in the original text.)

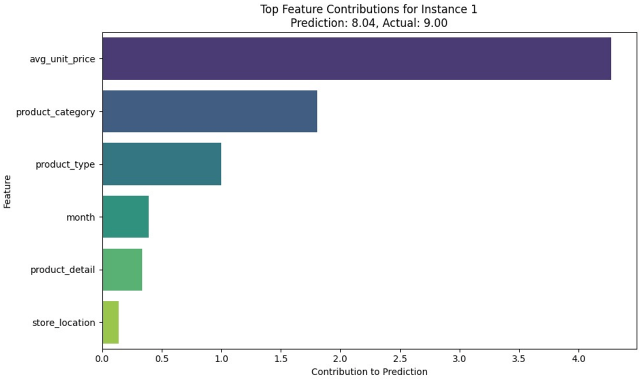

Fig. 2. Instance 1: Hell's Kitchen, Tea, Herbal Sales Prediction Factors.

Instance 2 - Lower Manhattan, Coffee, Cappuccino **Main characteristics:** Lower Manhattan location, Coffee product type (Cappuccino). **Contributions to prediction:** Location and product type dominate. **Explanation:** In business areas, cappuccino has stable sales throughout the year.

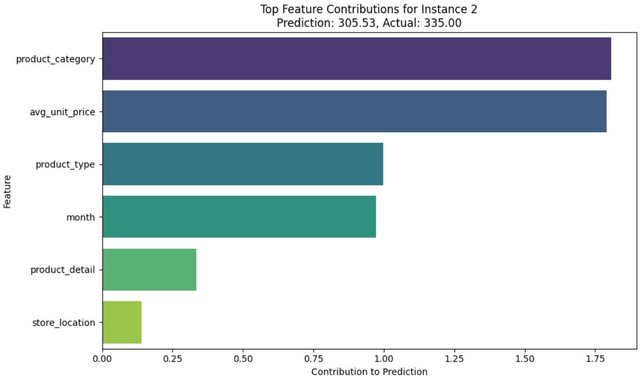

Fig. 3. Instance 2: Lower Manhattan, Coffee, Cappuccino Sales Prediction Factors.

Instance 3 - Astoria, Bakery, Croissant **Main characteristics:** Astoria location, Bakery product category, February month. **Contributions to prediction:** Product category and location are essential. **Explanation:** Pastries sell best in cold months, associated with hot drinks.

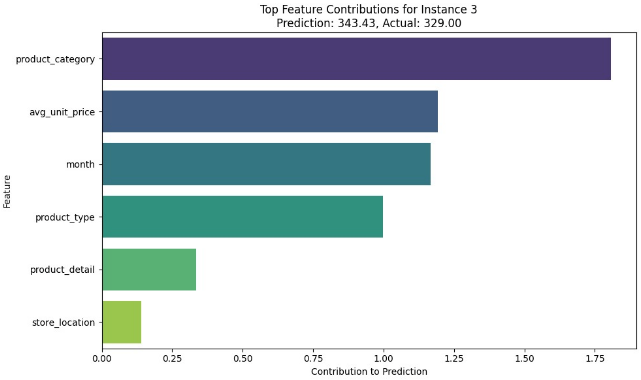

Fig. 4. Instance 3: Astoria, Bakery, Croissant Sales Prediction Factors.

Instance 4 - Hell's Kitchen, Coffee, Latte **Main characteristics:** Hell's Kitchen location, Coffee product type (Latte). **Contributions to prediction:** Product type and subtype are dominant. **Explanation:** Lattes are popular in this area, regardless of price.

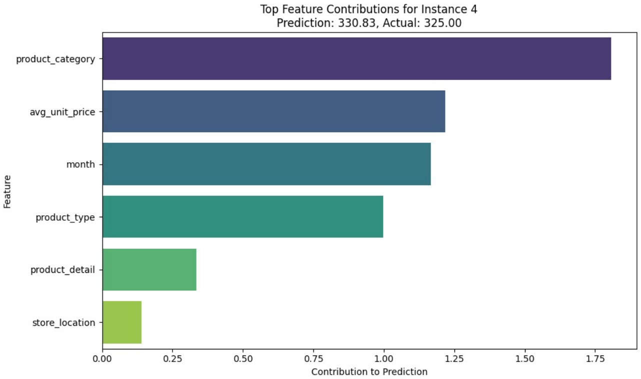

Fig. 5. Instance 4: Hell's Kitchen, Coffee, Latte Sales Prediction Factors.

6 Conclusions

This analysis project has successfully demonstrated the ability to analyze and model Maven Roasters' sales data to extract information and develop predictive models. The main objective of predicting the best-selling product per location was achieved by applying statistical methods and machine learning algorithms.

Efficiency of Predictive Models:
Random Forest proved to be the best performing model for predicting sales quantities, achieving an R^2 accuracy of 0.97, according to the final project evaluations. This model managed to efficiently capture complex and non-linear relationships between variables, such as seasonality and location-specific preferences.

Logistic Regression was successfully used to classify products into "popular" and "less popular," obtaining an accuracy of 0.91. This model offers valuable insight into the factors determining a product's popularity.

Linear Regression served as a baseline model, offering an initial perspective on linear relationships in the data, with an R^2 accuracy of 0.51.

In conclusion, this project has transformed raw transaction data into a set of actionable insights, capable of supporting the decision-making process and contributing to increasing the operational efficiency of Maven Roasters coffee shops.

Link to the database (coffe-shop-sales-revenue.csv[1]).
Link to the Git project (Coffe-Shop-sales[2]).

[1] https://www.kaggle.com/datasets/agungpambudi/trends-product-coffee-shop-sales-revenue-dataset
[2] https://github.com/PopLuks/Coffe-Shop-sales

References

1. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001)
2. Draper, N.R., Smith, H.: Applied Regression Analysis. 3rd edn. Wiley, New York (1998)
3. Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. 3rd edn. Wiley, Hoboken (2013)
4. Sitthiyot, T., Holasut, K.: A simple method for estimating the Lorenz curve. Humanities and Social Sciences Communications **8**(1), 1–10 (2021)
5. Quinlan, J.R.: Induction of Decision Trees. Machine Learning **1**(1), 81–106 (1986)
6. Rahm, E., Do, H.H.: Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin **23**(4), 3–13 (2000)
7. Friendly, M.: Corrgrams: Exploratory Displays for Correlation Matrices. The American Statistician **56**(4), 316–324 (2002)
8. Tan, J., Yang, J., Wu, S., Chen, G., Zhao, J.: A Critical Look at the Current Train/Test Split in Machine Learning. arXiv preprint arXiv:2103.09015 (2021)
9. Luengo, J., Garcia-Gil, D., Ramirez-Gallego, S., Garcia, S., Herrera, F.: Big Data Preprocessing: Enabling Smart Data. Springer, Cham (2020)
10. Nuzzo, R.L.: Histograms: A Useful Data Analysis Visualization. PM&R **11**(10), 1143–1146 (2019)