Walmart Grocery Price Prediction

Group Name: G

Group Members:

First name	Last Name	Student number
Dev	Makwana	C0885064
Joel	Crasto	C0883863
Krina	Patel	C0886861
Mahaveersinh	Chauhan	C0884854
Trushna	Patel	C0886910

Submission date: **15/08/2023**

Table of Contents

Abstract

The focus of the Walmart Price Prediction Project is the use of the Kaggle-sourced Walmart Grocery Product Dataset. The fundamental goal of this project is to develop an accurate and effective regression model for price prediction. The project includes a thorough procedure that includes data collecting, rigorous cleaning, and crucial preparation stages. Additionally, complex feature engineering takes place out, including the conversion of categorical data into vectors that have been one-hot encoded. Notably, the research effectively minimizes the dimensionality of the data using principal component analysis (PCA). The selection and assessment of a suitable regression model will be the outcome of the project work. The project intends to advance our knowledge of the changing nature of the grocery retail sector by offering insightful information about the variables that affect product prices at Walmart. The Walmart Price Prediction Project has greater significance than just its technological components. The project provides useful insights into the elements that affect product pricing at Walmart by probing the dynamics of the food retail industry. This information can significantly affect the retail sector and consumers by illuminating patterns, market dynamics, and other factors that affect price variations.

Introduction

The primary lessons and conclusions from the analysis we conducted for Walmart are included in this report. Walmart is considered as one of the leading retailers in the entire world, so it is very important for them to have accurate price predictors. Every department's sale can be impacted by a variety of factors; therefore, it is important to identify the relative ones that drive sales and utilize them to build a model that can aid in somewhat accurate pricing forecasting.

For this project we have used the “Walmart Product” dataset which is openly available on Kaggle. In this dataset it has multiple columns such as product size, product category, shipping location and many more. Using all these features we need to create multiple regression models which can predict the accurate price. Root Mean Square Error (RMSE) to evaluate the result of the model.

The main goal of this project is to analyze the grocery data from Walmart to predict the price of different groceries.

During the project work the following questions were answered:

- What is the average price of products in Walmart Grocery?
- What is the average product size (in terms of weight) in Walmart Grocery?
- What are the most popular brands in Walmart Grocery?
- What are the most popular products in Walmart Grocery?
- Predict the price of the products?
- Check department-wise sales at stores and for different stores of Walmart?

Methodology

The goal of the project is to predict the price of the product based on the analysis of the features in the dataset. The basic steps implemented for creating the project is outlined below:

1. Dataset details

The dataset used in this assignment is from Kaggle (Jeff, 2022) platform that consists of Walmart Product. The dataset contains 568534 recorded products. Table 1 describes features and their meaning for analyzing and predicting.

Field Name	Description	Example
SHIPPING_LOCATION	The location where the product is shipped from. (String)	79936
DEPARTMENT	The department in which the product is categorized. (String)	Deli
CATEGORY	The category in which the product is categorized. (String)	Hummus, Dips, & Salsa
SUBCATEGORY	The subcategory in which the product is categorized. (String)	White Wine
BREADCRUMBS	The breadcrumbs for the product. (String)	Deli/Hummus, Dips, & Salsa
SKU	The SKU for the product. (String)	110895339
PRODUCT_URL	The URL for the product. (String)	https://www.walmart.com/ip/Marketside-Roasted-Red-Pepper-Hummus-10-Oz/110895339?fulfillmentIntent=Pi...
PRODUCT_NAME	The name of the product. (String)	Marketside Roasted Red Pepper Hummus, 10 Oz
BRAND	The brand of the product. (String)	Marketside
PRICE_RETAIL	The retail price of the product. (Float)	2.67
PRICE_CURRENT	The current price of the product. (Float)	2.67

Field Name	Description	Example
PRODUCT_SIZE	The size of the product. (String)	10
PROMOTION	The promotion for the product. (String)	NULL
RunDate	The date on which the data was collected. (Date)	2022-09-11 21:20:04
Tid	Transaction ID	16163804

2. Analysis Process

I. Data Cleaning

- Our dataset contains few null values so that may lead to wrong prediction of the price. To prevent this and as per our mutual discussion we have decided to remove all the null values.
- There are some redundant columns such as RunDate, SKU, Promotion, tid, Product URL, and Subcategory. Before proceeding further, we have removed those columns, so it won't affect our result.
- The feature called Product Size was of type object which needs to be converted to numeric form. Thus, we use to_numeric() method of pandas to convert. This resulted in many null values in the same column which were dropped.

II. Exploratory Data Analysis

Exploratory Data Analysis is used to get insight of the dataset. Some visualized questions can be answered using a graphical representation of the dataset.

1) **Average Price of Products by Department:** The prices of Alcohol drink is more than other products in the grocery store. Hence, it shows the highest bar in the chart.

Figure 1 Average Price of Products By Department

2) Average Product Size by Department:

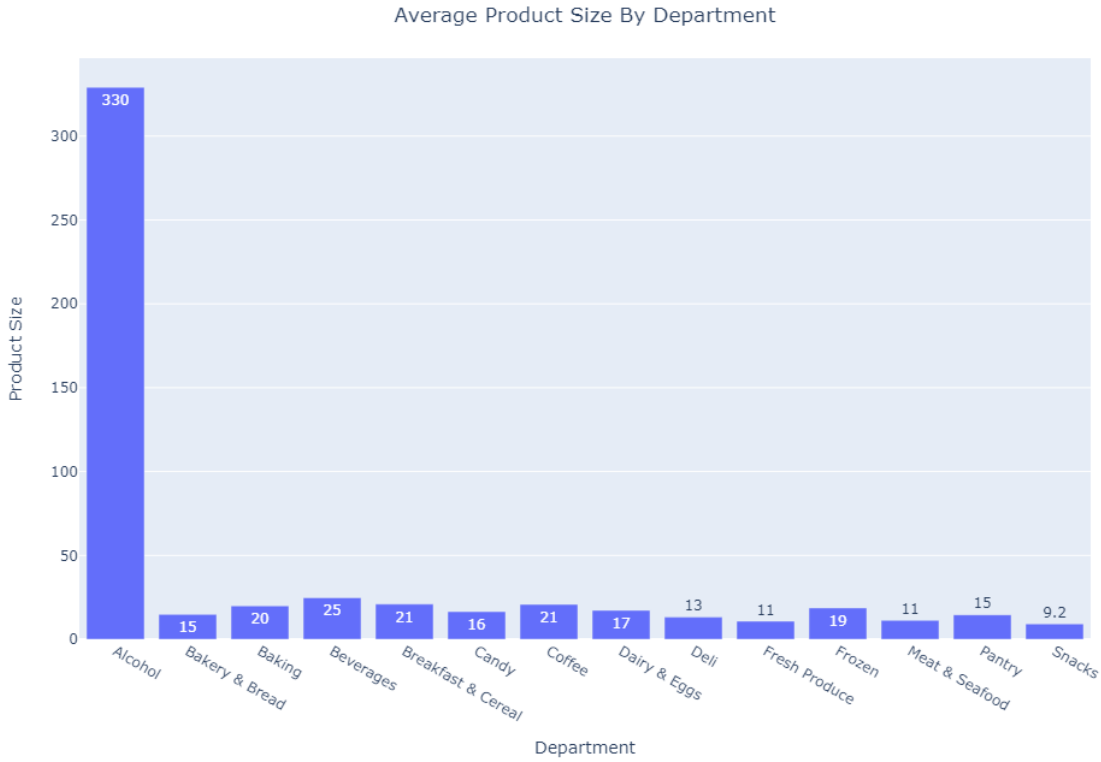

Figure 2 Average Product Size by Department

3) **Top 10 Popular Brands:** It is clearly evident from the below pie chart that the Great Value brand is the most popular in Walmart Grocery.

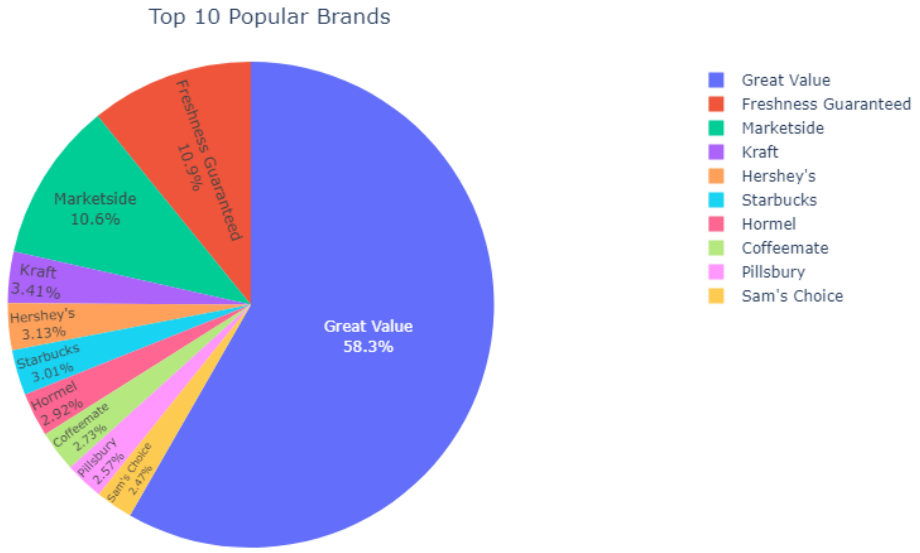

Figure 3 Top 10 Popular brands

4) **Top 20 Popular Products:** In the list of top 20 popular brands, it is observed that the products of Great Value Brands are more popular. We can also infer that chocolate products are also popular.

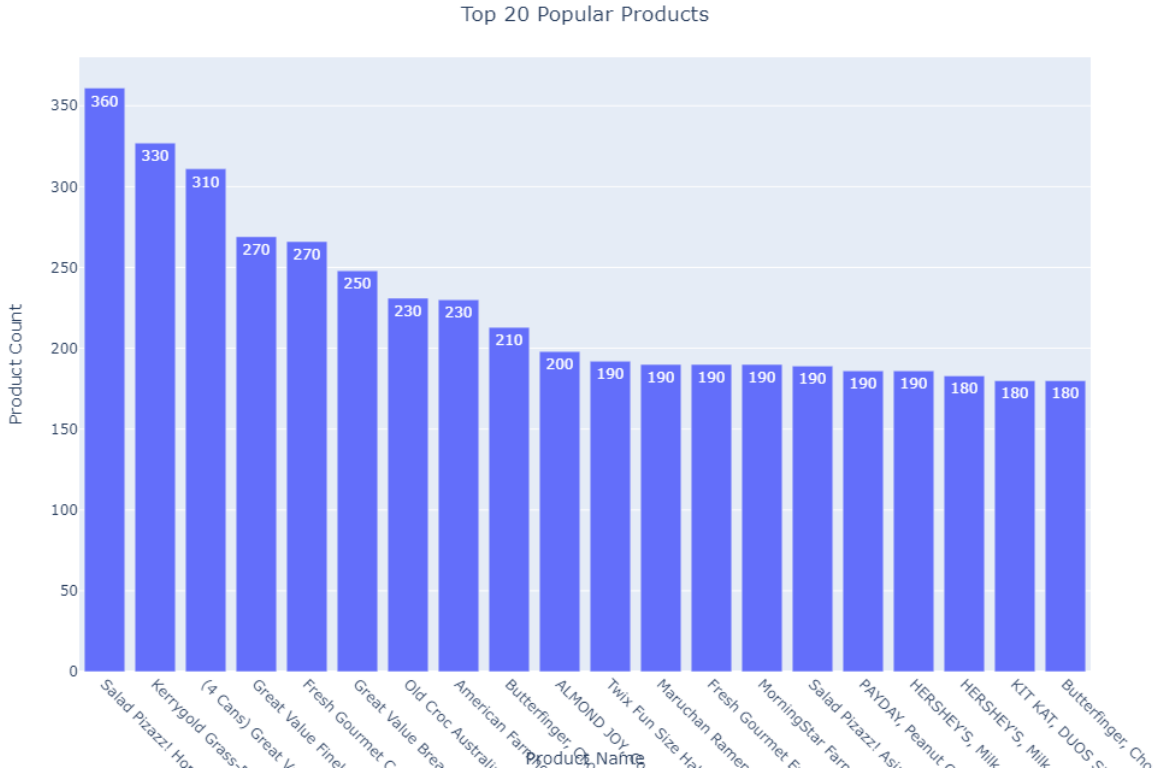

Figure 4 Top 20 Popular Products

5) **Total Price by Shipping Location corresponding to different Departments:** Following stacked bar graph represents the total of all prices at different shipping locations department wise. It is observed that the Baking and Breakfast and Cereal Department shows the highest prices of products while Fresh Produce and Meat and Seafood shows the lowest product prices.

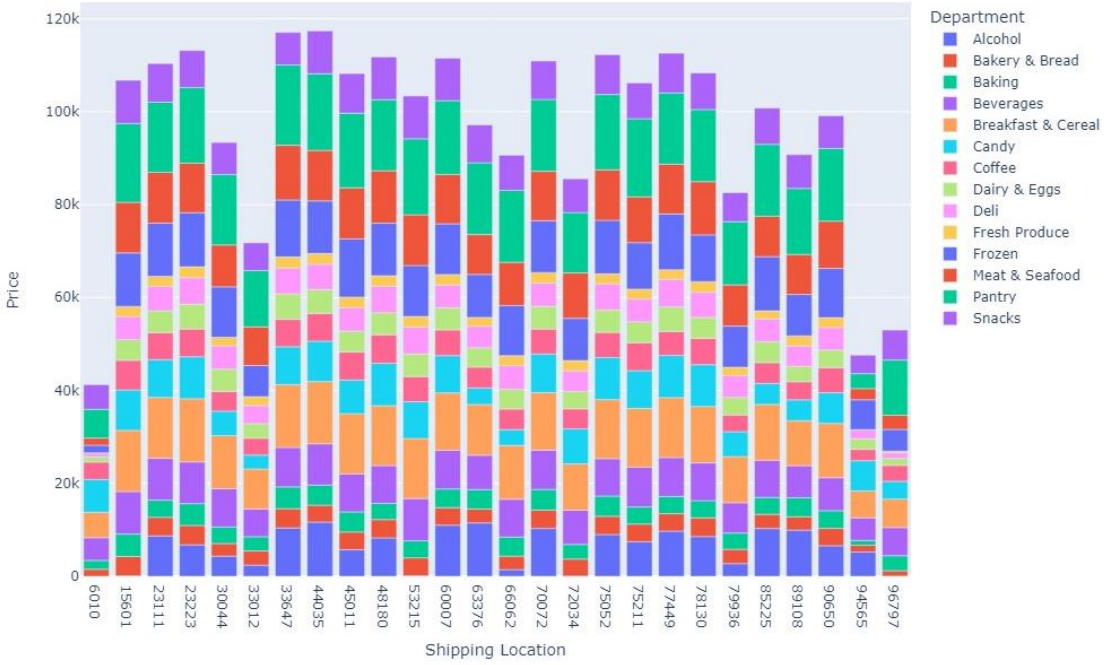

Figure 5 Total Price by Shipping Location corresponding to different Departments

6) **Percentage of Product counts by Department:** The pie chart below shows the percentage count of number of products with respect to various departments. The Pantry and Breakfast and Cereals department has the highest products while Fresh Produce and Alcohol had lesser products.

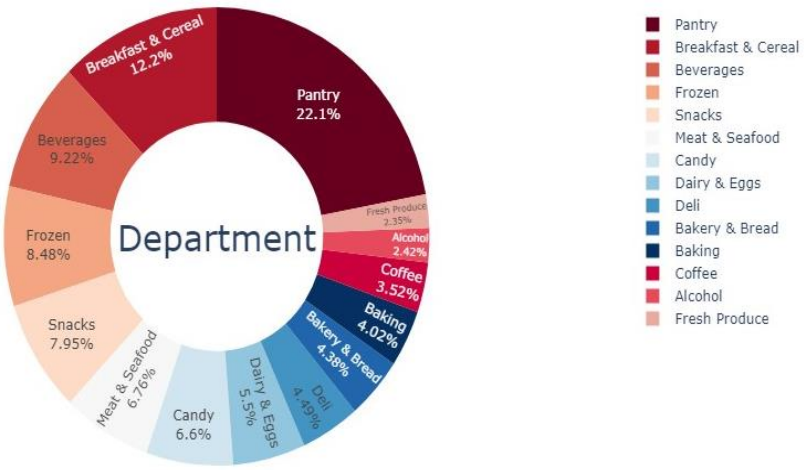

Figure 6 Percentage of Product counts by Department

7) **Total Price of Top 20 Categories:** The Meat and Seafood section observed the highest sale of around 200k followed by the Breakfast and Cereals department which made around 140k.

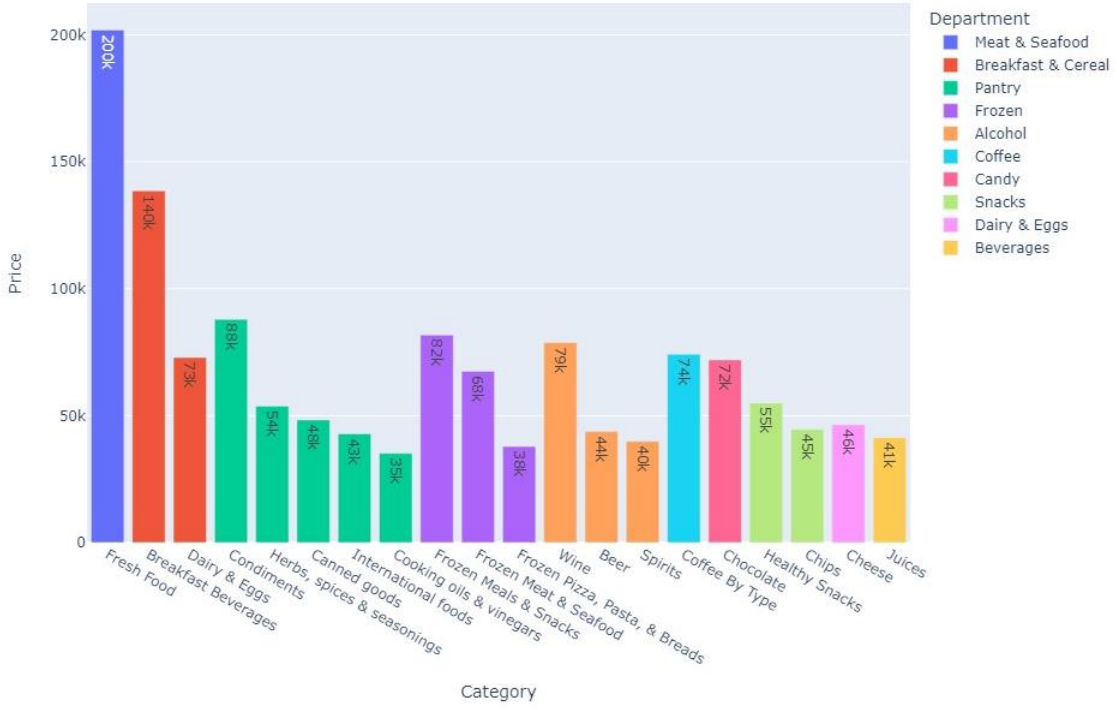

Figure 7 Total Price of Top 20 Categories

8) **Average Price of Top 20 Brands:** Below chart shows the top 20 selling brands of all products.

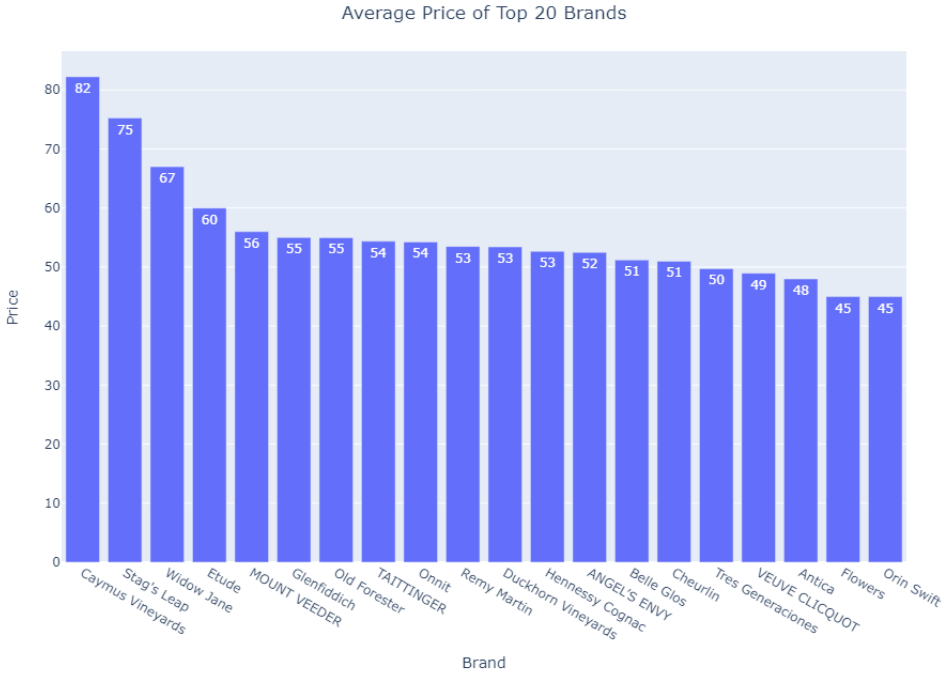

Figure 8 Average Price of Top 20 Brands

9) **Price Distribution Count by Department:** The Graph says that the price distribution is right skewed. It is not normally distributed, so we need to convert it to normal distribution.

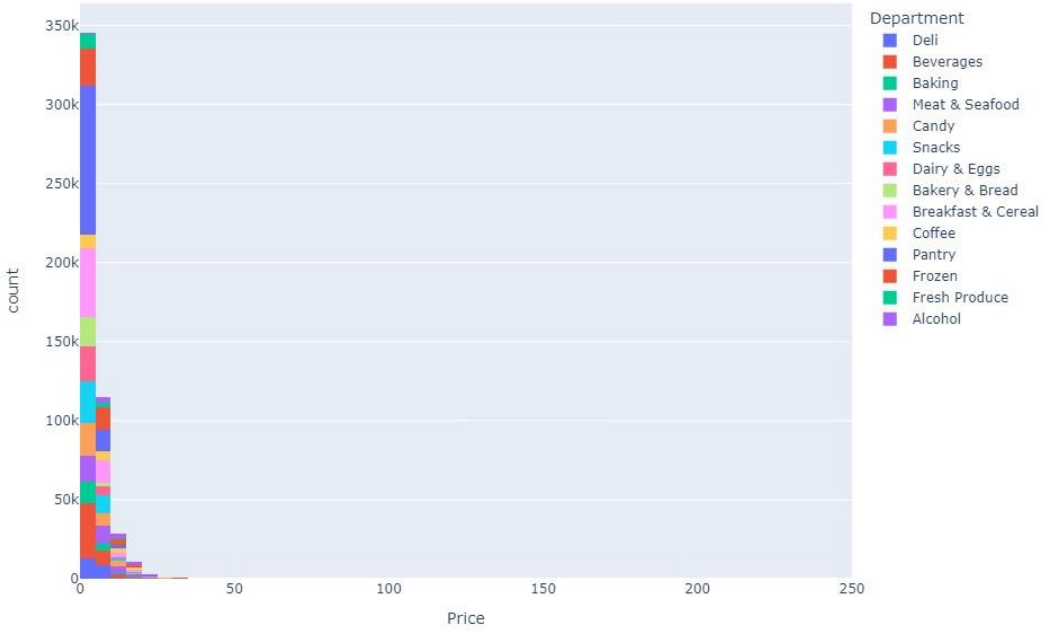

Figure 9 Price Distribution count by Department

10) **Logarithmic Price Distribution by Department:** To convert the price into normal distribution we have applied a log transformation function.

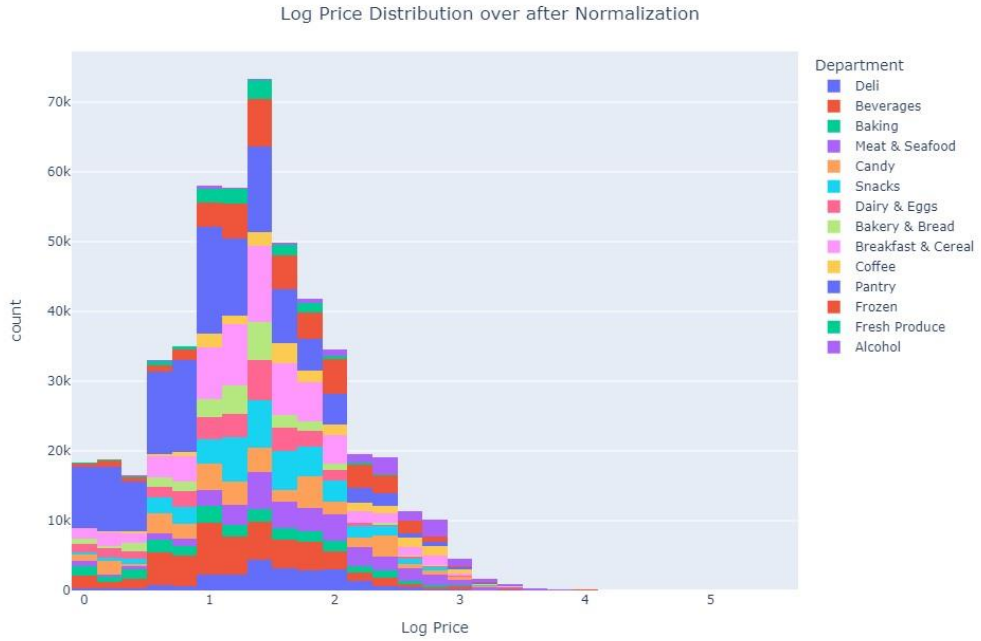

Figure 10 Logarithmic Price Distribution by Department

III. Feature Engineering

- After cleaning the whole data, for the prediction still there are few unrelated columns like shipping location, breadcrumbs, product name and price current which needs to be removed.
- Department, Category and Brand columns are categorical so those need to be converted into numerical columns using the get_dummies method.

IV. Principal Component Analysis

- PCA (Principal Component Analysis) is a method to reduce the number of features by maintaining the significant variance of the data.
- The features such as Department, Category, and Brand are categorical.
- After applying the get_dummies method there were more than 4000 columns so after applying the PCA method the resulting features were 25.

V. Machine Learning Model

To assess the model performance on fresh data, we employ a train-test split in machine learning. The dataset will be split into a training set and a test set. The test set assesses the model's correctness or error on fresh data, while the training set is used to fit the model.

Linear Regression: Linear regression (*About Linear Regression | IBM*, n.d.) analysis is used to predict a variable's value based on another variable's value. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Lasso Regression: In statistics and machine learning, lasso (least absolute shrinkage and selection operator; also Lasso or LASSO) (*Lasso (Statistics) - Wikipedia*, 2022) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the resulting statistical model.

Decision Tree: Decision Tree (*Python | Decision Tree Regression Using Sklearn - GeeksforGeeks*, 2018) is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables.

CatBoost Regression:CatBoost (*How CatBoost Algorithm Works&Mdash;ArcGIS Pro | Documentation*, n.d.) is a supervised machine learning method that is used by the Train

Using AutoML tool and uses decision trees for classification and regression. As its name suggests, CatBoost has two main features, it works with categorical data (the Cat) and it uses gradient boosting (the Boost).

XGBoost Regression: XGBoost (*How XGBoost Works - Amazon SageMaker*, n.d.) is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models**.**

Random Forest Regression: A random forest (*Sklearn.Ensemble.RandomForestRegressor*, n.d.) is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

Results

The Machine Learning Models listed in the previous chapter were implemented to analyze the results on the decided dataset using the measure of Training and Testing scores.

Algorithms	Training R2 Score (%)	Testing R2 Score (%)
Linear Regression	21.56	21.46
Lasso Regression	21.55	21.46
Decision Tree	91.39	90.40

The Linear and Lasso Regression provides accuracy similar results of 21%. The reason for this might be because, they both does not handle non-linearity and are not robust to outliers. Despite being simple decision tree regressor, it performed better than other regression models mentioned above. The accuracy provided by Decision Tree is 90 % on both training and testing sets. Thus, the results are more precise for ensemble techniques, they are explored further.

Tree-based Models	Training R2 Score (%)	Testing R2 Score (%)
CatBoost	74.18	74.73
XGBoost	75.60	75.08
Random Forest	90.40	91.39

The Cat Boost Regressor is giving the same testing and training, 74 % accuracy which indicates model is fitting very well and does not overfit. Also, XG Boost Regressor is like Cat Boost Regressor but varies by only 1%. Thus, the RandomizedSearchCV was implemented to get best parameters by tuning some hyper-parameters including n_estimators, max_depth, learning_rate, and so on. With these features, XGBoost model performs the best with a training and testing score of 86%.

Random Forest Regressor performs the best amongst all the other models on both training and testing data. The accuracy is 89% and 90% for training and testing respectively. Although the Randomized Search CV was applied, Random Forest was not able to give better accuracy than it was before tuning hyper-parameters.

Conclusion and Future Work

This project aims to contribute to the field of prediction by leveraging Machine Learning techniques to develop a model for Walmart Grocery Price prediction. By accurately predicting the outcomes of Grocery prices, the developed model will have significant implications for customers, product sellers and Walmart. The predicted insights provided by this system can help retailers like Walmart allocate resources more effectively and manage their inventories more effectively. The results were spectacular for Random Forest with training and testing scores of 89% and 90%, respectively.

There are great prospects of enhancing the Walmart grocery price prediction system in the coming years. We can improve our pricing forecasts' efficiency and ability to respond to changes by utilizing more sophisticated algorithms and real-time data. We might also make personalized purchasing suggestions and collaborate with reward programs to offer exclusive discounts. It might also be possible for Walmart to improve its planning by investigating issues like how price-sensitive consumers are and projecting what products may be in demand.

References

Walmart Products. (n.d.). Walmart Products | Kaggle. https:///datasets/thedevastator/product-prices-and-sizes-from-walmart-grocery

Python | Decision Tree Regression using sklearn - GeeksforGeeks. (2018, October 4). GeeksforGeeks. https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/

How XGBoost Works - Amazon SageMaker. (n.d.). How XGBoost Works - Amazon SageMaker. https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

sklearn.ensemble.RandomForestRegressor. (n.d.). Scikit-learn. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

How CatBoost algorithm works—ArcGIS Pro | Documentation. (n.d.). How CatBoost Algorithm Works&Mdash;ArcGIS Pro | Documentation. https://pro.arcgis.com/en/pro-app/latest/tool-reference/geoai/how-catboost-works.htm#:~:text=CatBoost%20is%20a%20supervised%20machine,gradient%20boosting%20(the%20Boost).

Lasso (statistics) - Wikipedia. (2022, August 22). Lasso (Statistics) - Wikipedia. https://en.wikipedia.org/wiki/Lasso_(statistics)#:~:text=In%20statistics%20and%20machine%20learning,of%20the%20resulting%20statistical%20model.

About Linear Regression | IBM. (n.d.). About Linear Regression | IBM. https://www.ibm.com/topics/linear-regression