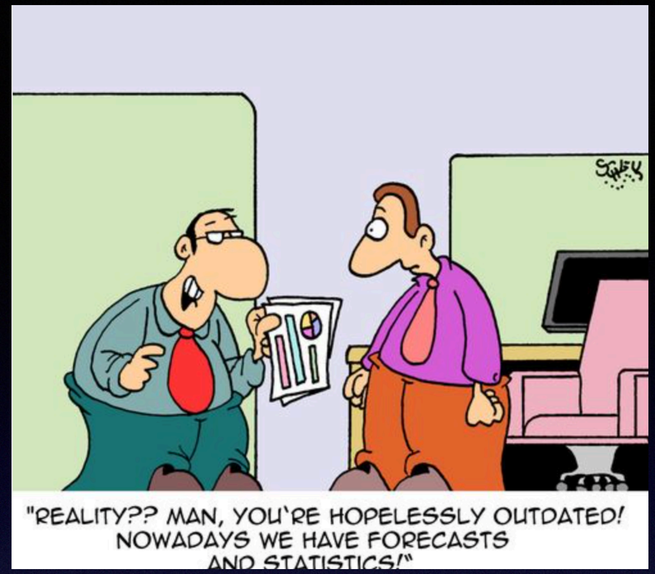


Capstone_Presentation

Demand Forecast Project

Presented by - Swati Sharma

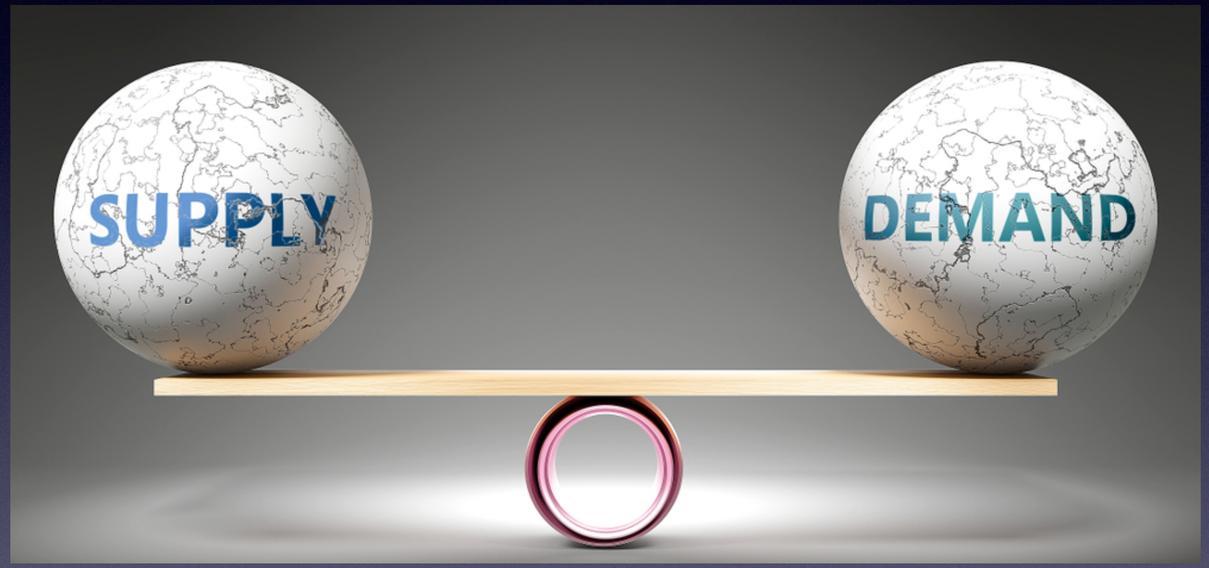
Context

Fresh foods, a challenge to any grocer or retailer

- Short shelf life,
- Appearance,
- Season dependency,
- Specific transportation,
- Storage requirements

Led to Demand and Supply Imbalance.

Food wastage, - 2.5 billion tons every year



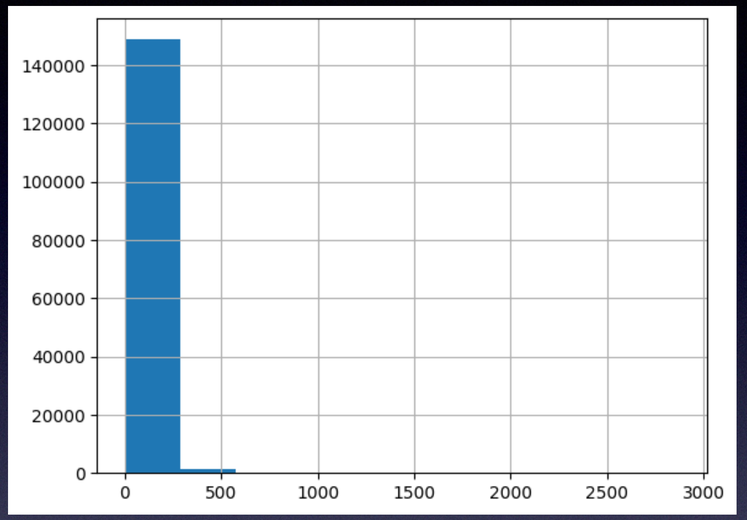
Need for a Demand Forecast system to resolve this issue at a global level.

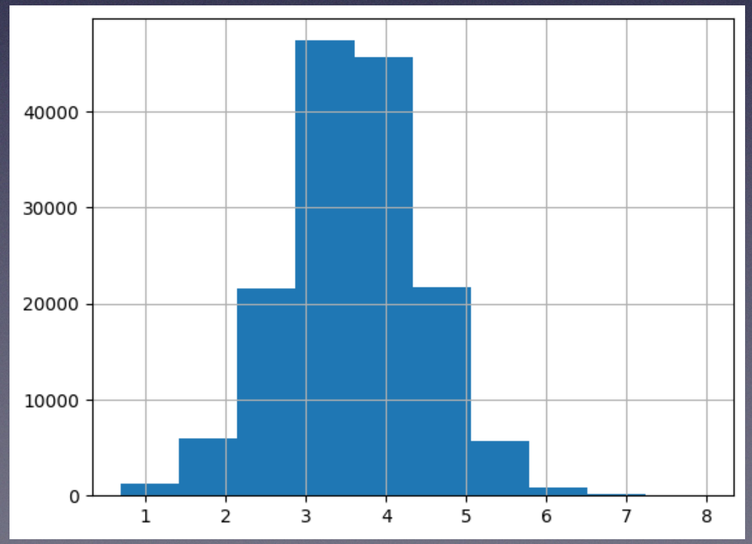
What is the Problem?

What opportunities exist for grocers or retailers to effectively develop and implement a new “Forecasting ML model” to better forecast the demand for perishable goods.

Data Wrangling

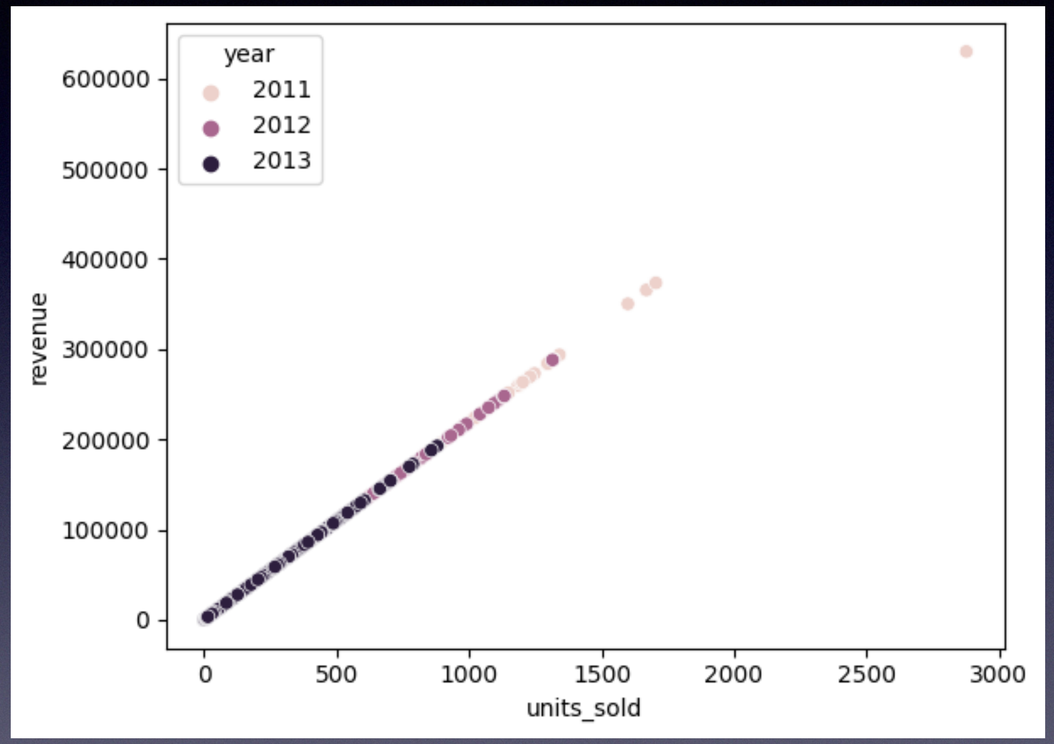
- Performed a series of processes to explore, and validate raw dataset
- Transformed into a high-quality and reliable data
- Normalized the target variable, units_sold - Smooth effect of Outlier and make data more uniform and easier to model.

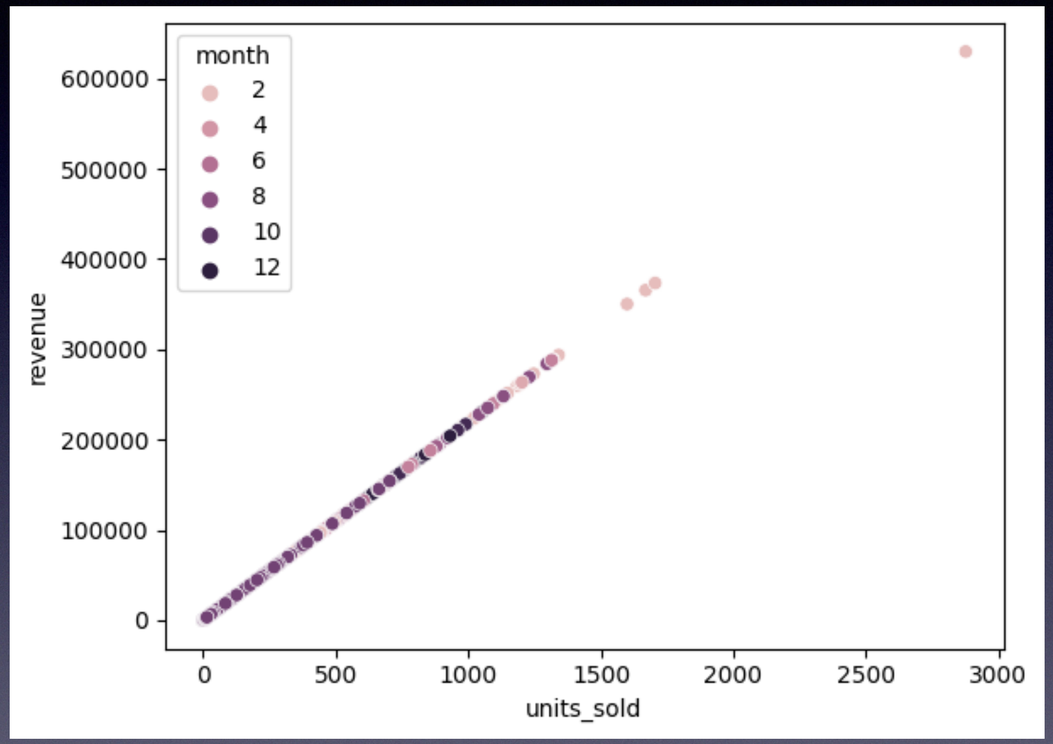




Data Source: Kaggle

Data Wrangling - Continue



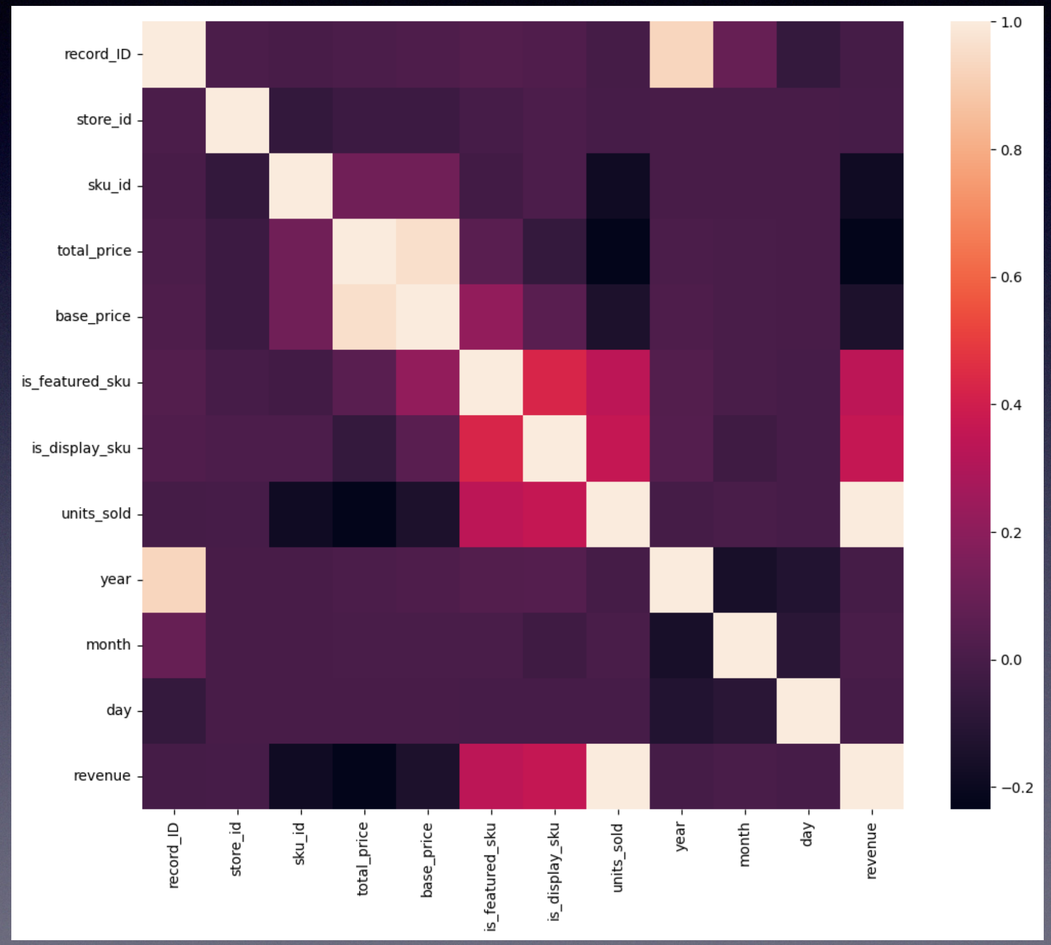


Decline in the sales per units sold between 2011 and 2013
February (Month 2) saw the maximum purchases
August (Month 8) saw the least amount of sales

Exploratory Data Analysis

Performed **Principal Component Analysis (PCA)** to reduce the scope of the feature space and identify the principal components.

Feature correlation heatmap - Gain a high level view of relationships amongst the features

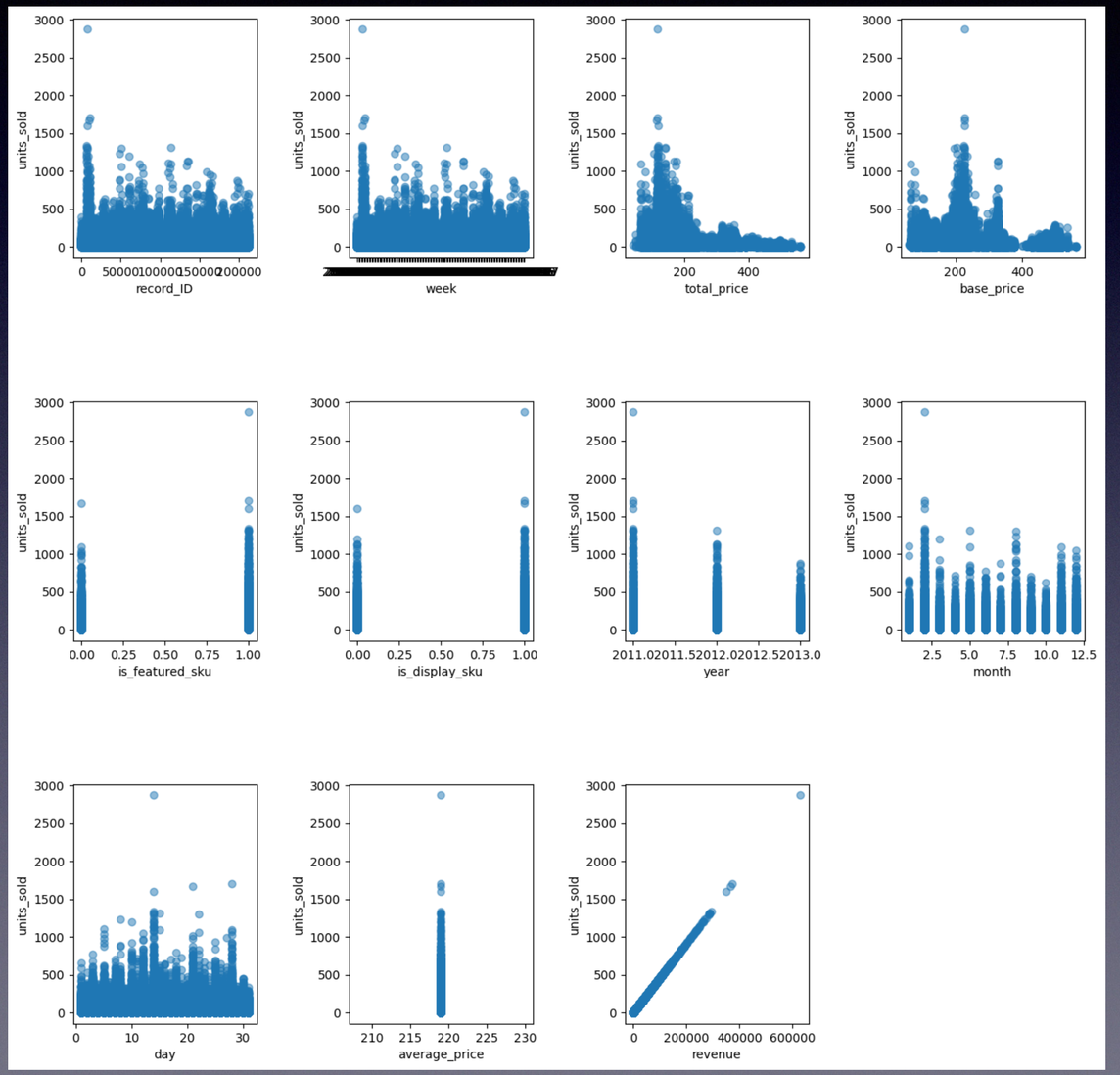


- There’s a strong positive correlation between units_sold and revenue.
- record_id and year are positively correlated.
- Strong positive correlation between sku_id and total_price and base_price.

Exploratory Data Analysis - Continue

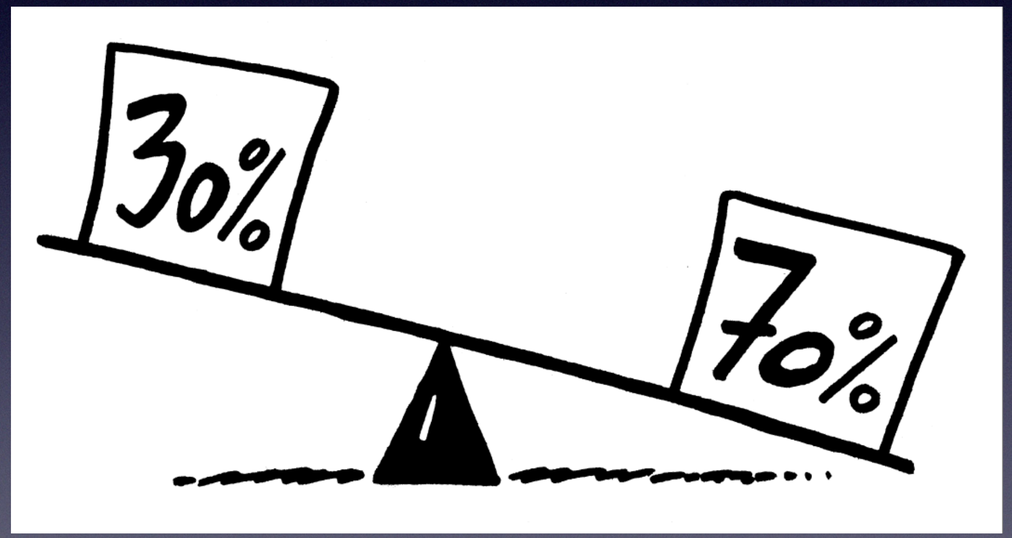
Create Scatter plots for visualizing the relationship between a numeric feature against target variable, units sold.

- Year, Month and Day seems useful to understand the seasonality and behavioral aspect.
- is_featured_sku and is_display_sku appear quite similar.
- There are some outliers present in almost all columns.



Further Analysis via Pre-processing and Training data

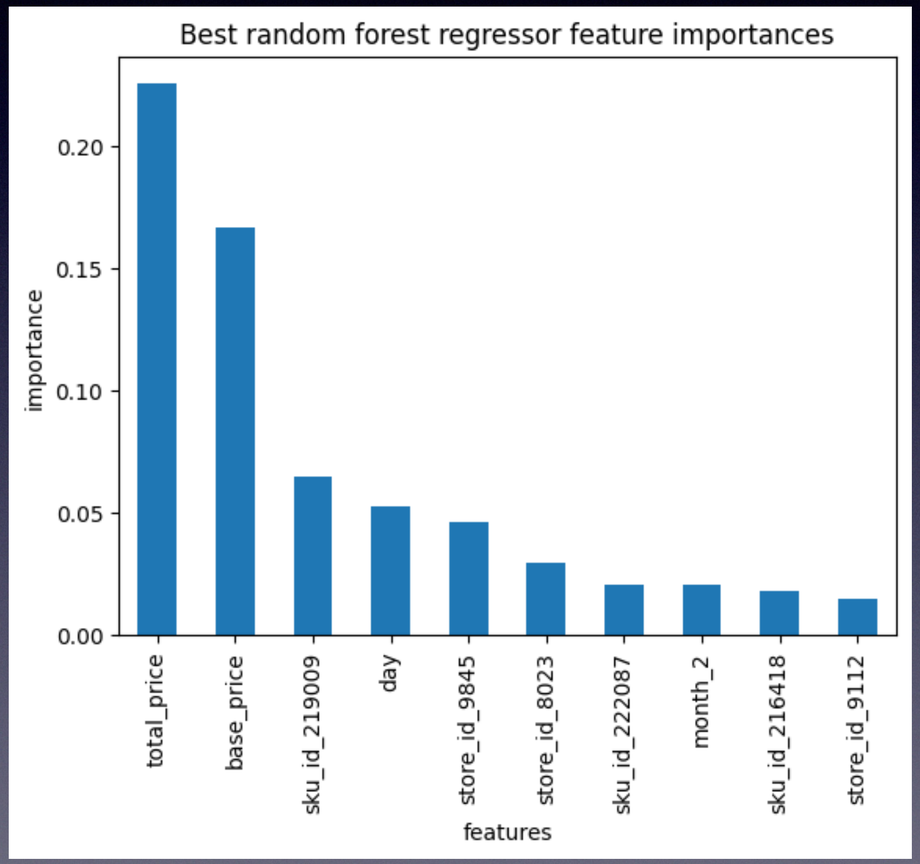
- Drop insignificant columns such as week, is_featured_sku, is_display_sku, average_price, revenue.
- Impute missing values.
- Converted categorical column “month”, "store_id", "sku_id", “year” into a format suitable for machine learning algorithms via “One-hot encoding”.
- Split the data into 70/30 train and test dataset.



Modeling Results and Analysis

Identified dominant features using Random forest regression model

- total_price
- base_price
- sku_id_219009
- day
- store_id_9845
- store_id_8023
- sku_id_222087
- month_2
- sku_id_216418
- store_id_9112

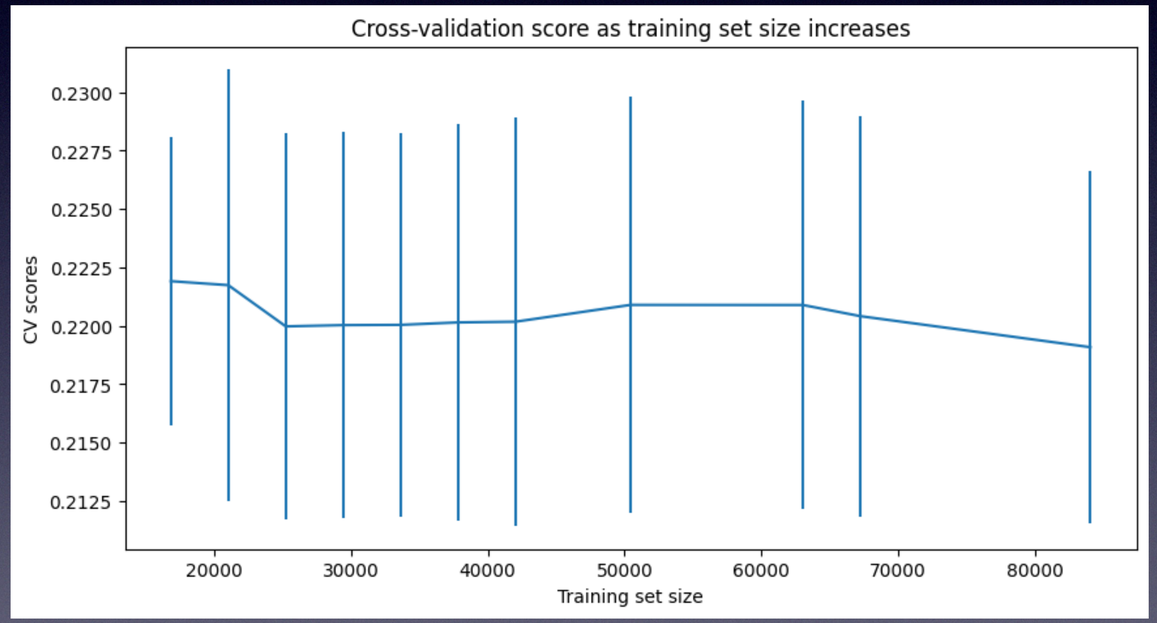


Model Evaluation: The random forest model has a lower cross-validation with least mean absolute error of 13.45
Exhibits less variability. Robust to outliers too.

Data quantity assessment

Per model CV score, the model performance quite leveled off.

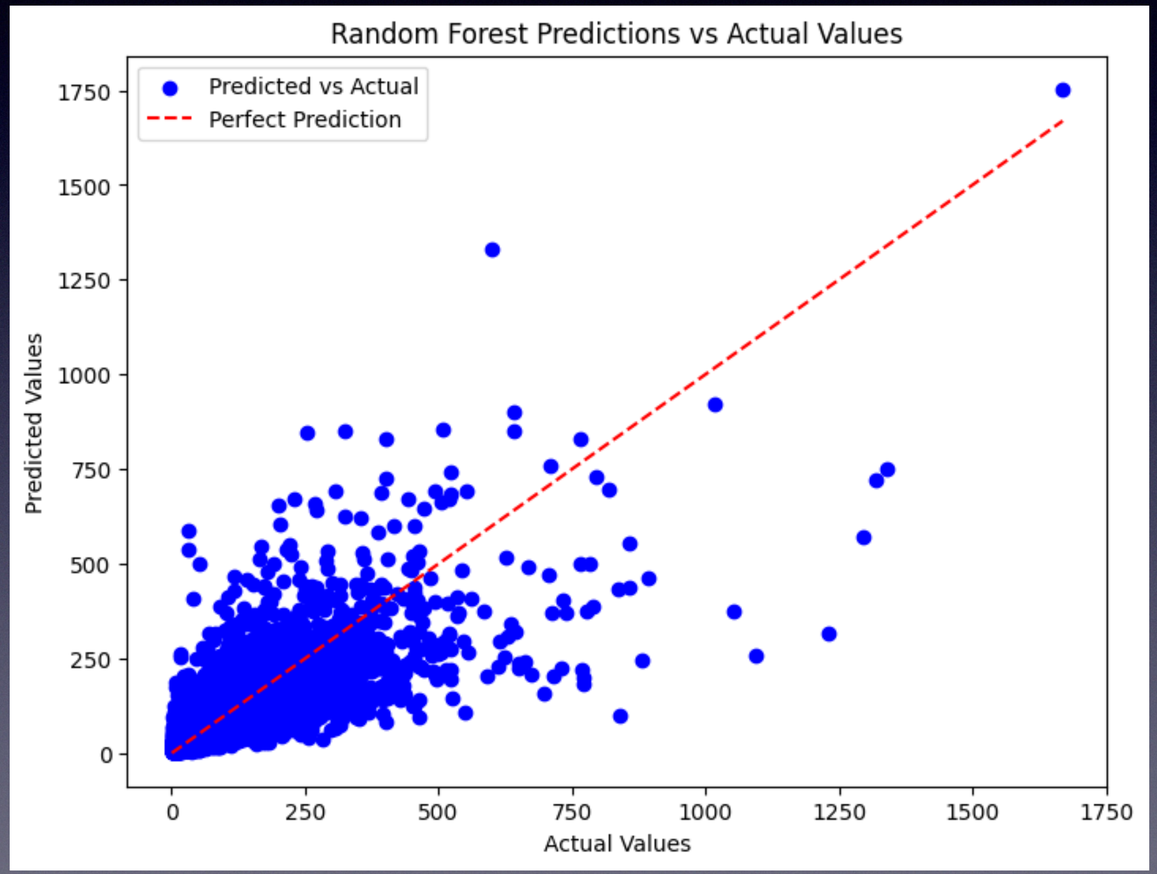
Saw an increase initially in model performance which then reduced slightly, then stayed same.



Visualize the Predictions

Plotted the predicted values against the actual values (y_test).

- Each point on the scatterplot represents a pair of **actual** and **predicted** values.
- The **red dashed line** represents perfect predictions (y = x). If all the points lie on this line, the model has made perfect predictions.



Recommendation and Key Findings

- Add more relevant features like external factors (economic indicators, holidays, etc.) to improve predictions.
- Determine specific products belong to sku id’s, 219009, 222087 and 216418 to predict future demand accurately.
- Research economical and environmental factors highlighting store id’s 9845, 8023, 9112.
- Include external data sources such as marketing campaigns, competitor actions to stay updated.
- Tune the model periodically with fresh data to keep it up to date.