Demand Forecast Project Report

Introduction

For any grocer or retailer, fresh foods poses a biggest challenge because of their short shelf life, appearance, season dependency, specific transportation and storage requirements and various other factors that can not just impact demand but also led to huge amount of food wastage. Globally, around 2.5 billion tons of food got waste every year, with US tops at first position with 60 million tons - 120 billion pounds food wastage every year. There is an extraordinary need for a better system to resolve this issue at a global level.

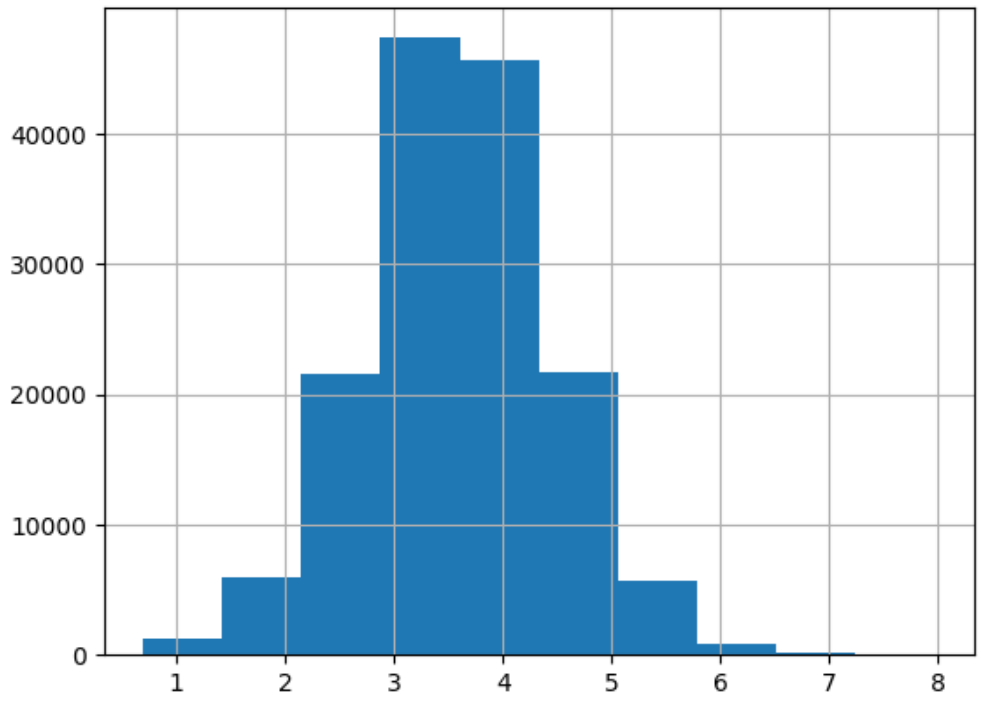
Problem Identification

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.

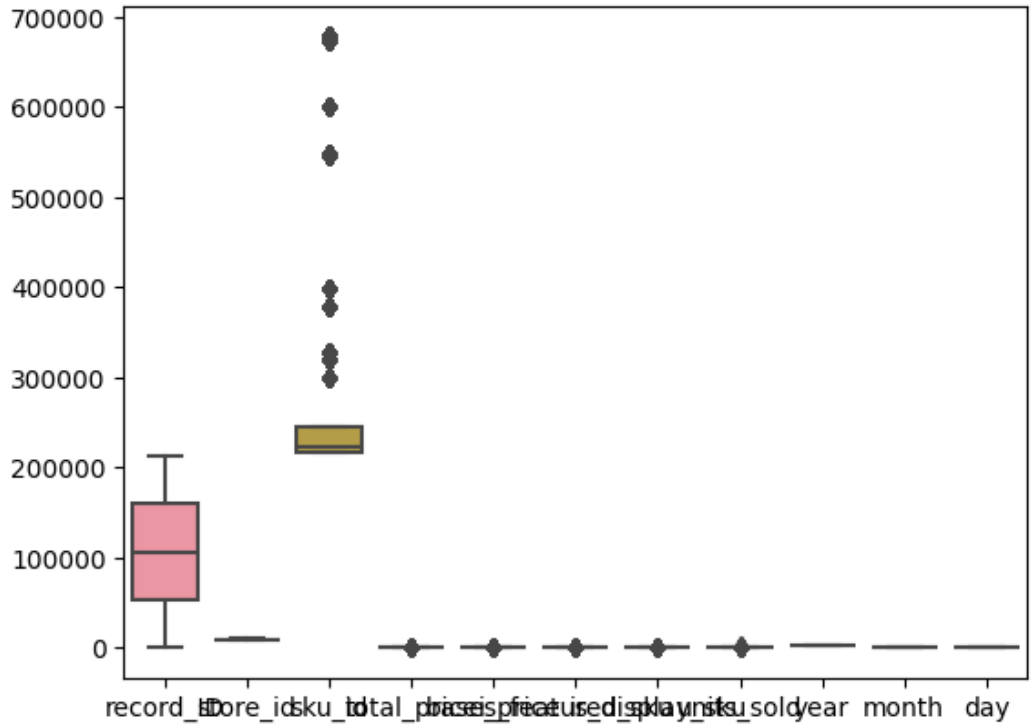
Data Wrangling

In this step, we performed a series of processes to explore, transform, and validate raw dataset retrieved into a high-quality and reliable data.

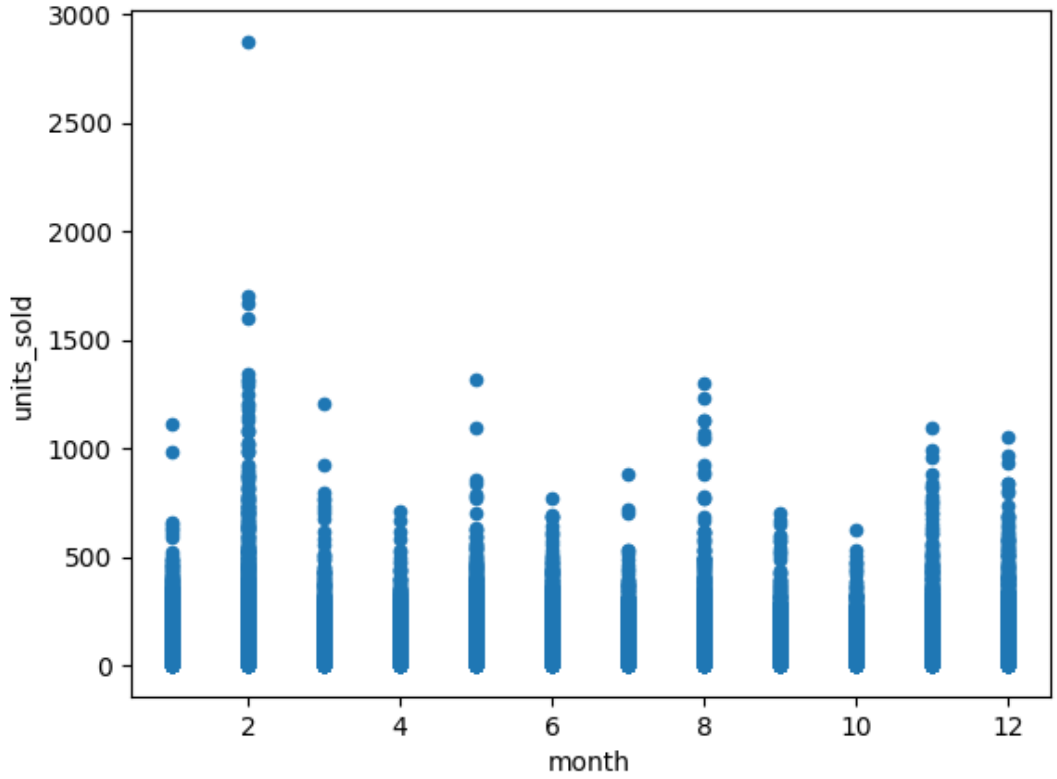
During our initial step, we first normalized the data, since there were some outliers in the column “units_sold”.



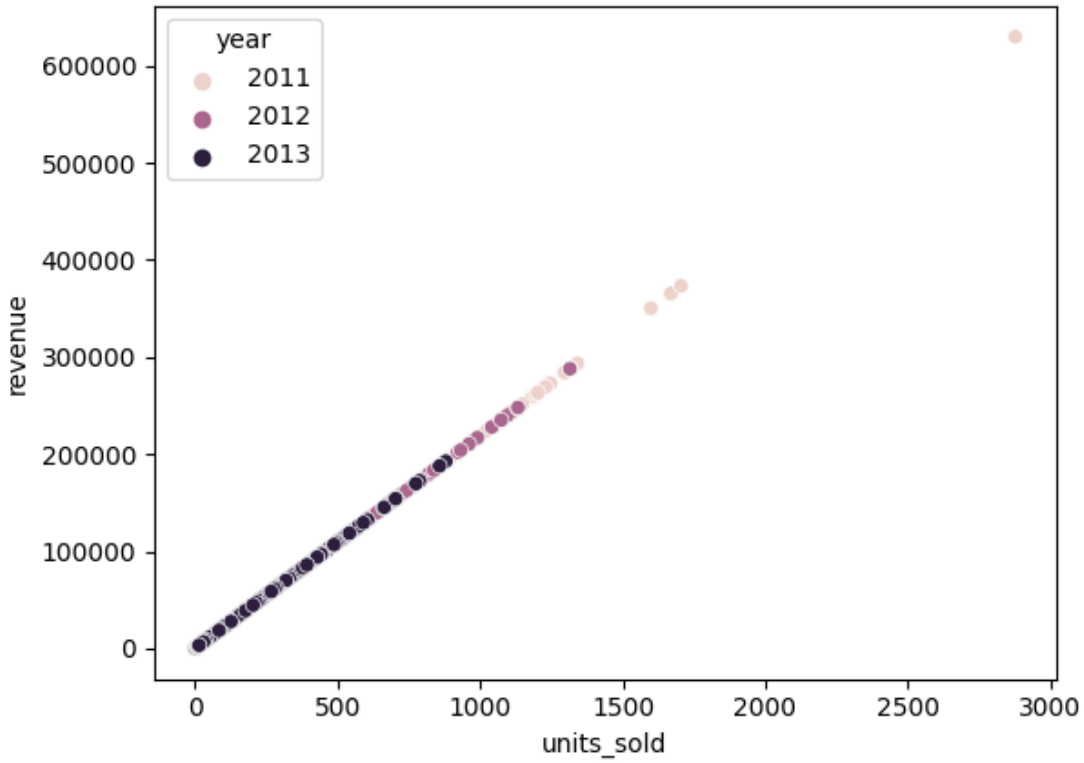
The next step was to create a seaborn boxplot to confirm for any outliers in any columns present.

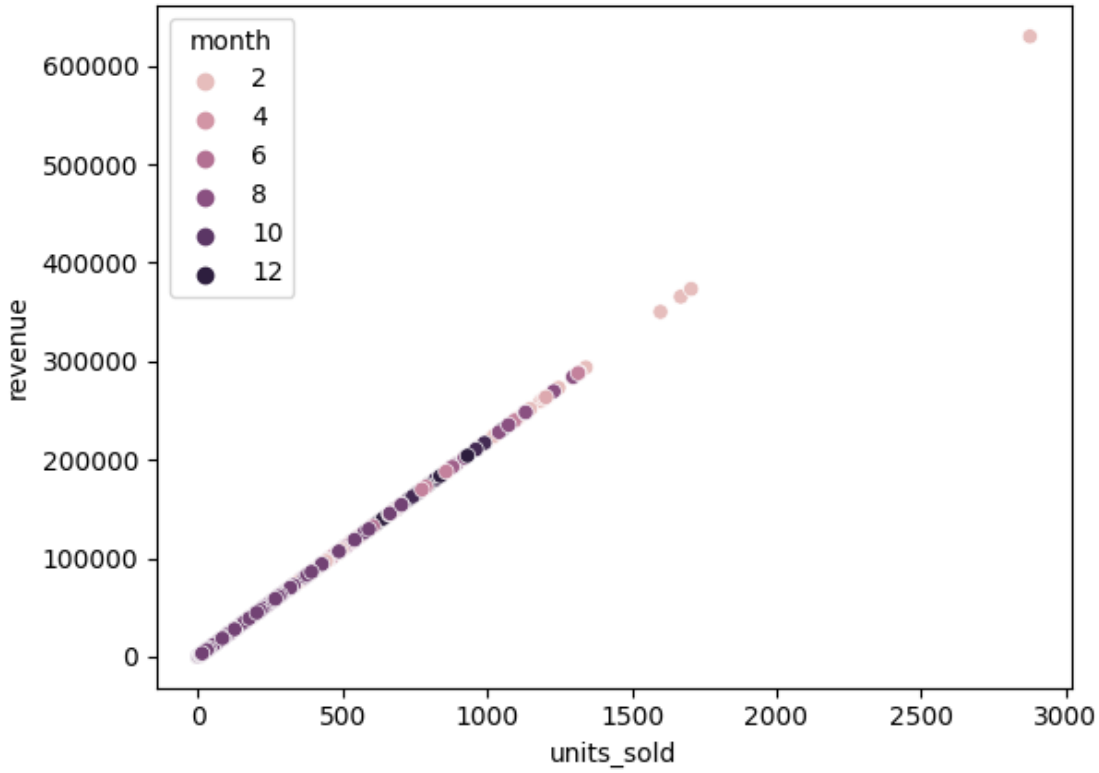


Later, we check the relationship between two columns, units_sold and months.



Furthermore, since our target feature is “units_sold”, we explored the relationship between year, month and units_sold and check if there's any trends between them.

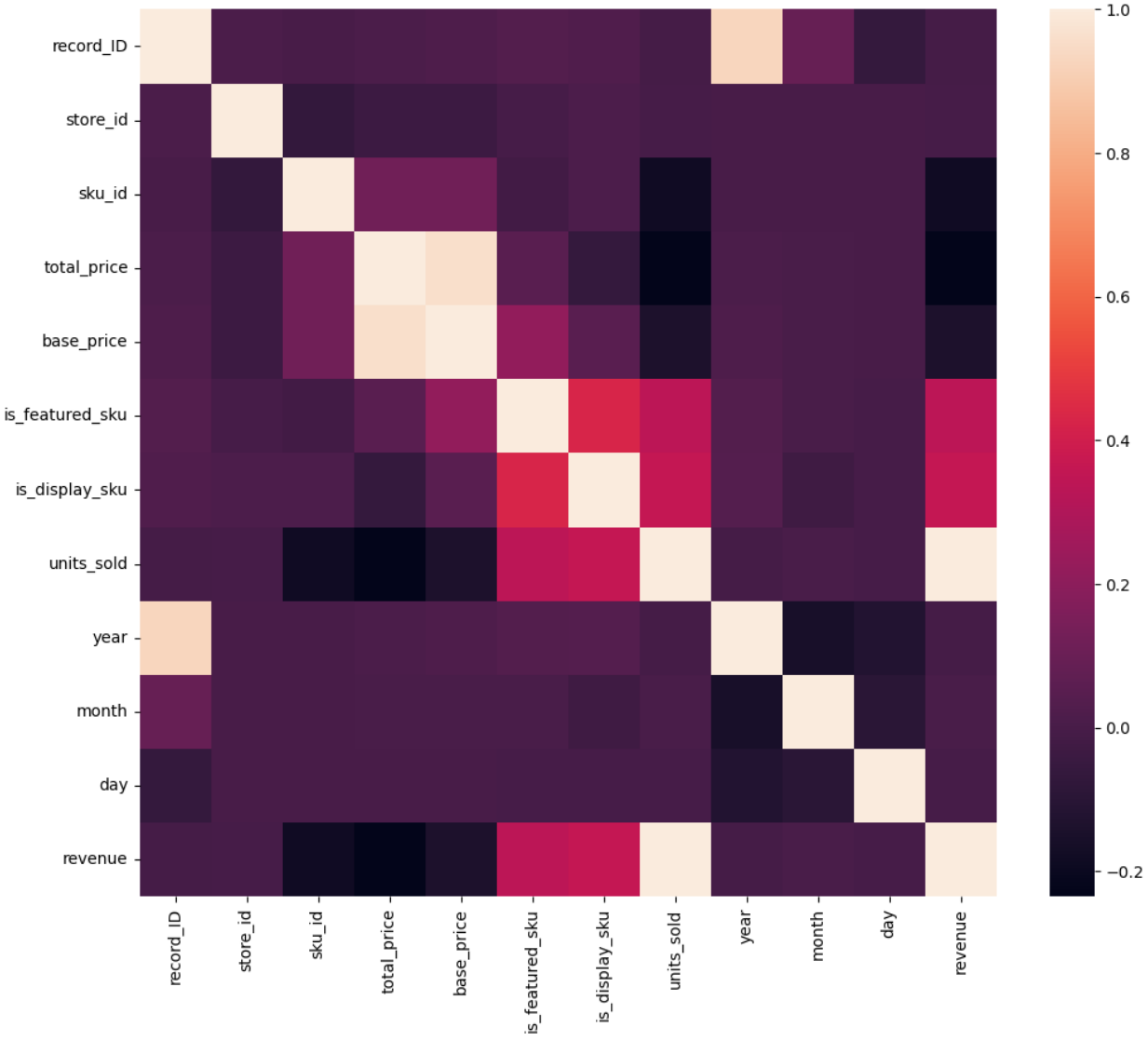




This indicates there's a decline in the sales/units sold from 2011 to 2013 and February has seen the maximum purchases whereas August is the month with least amount of sales.

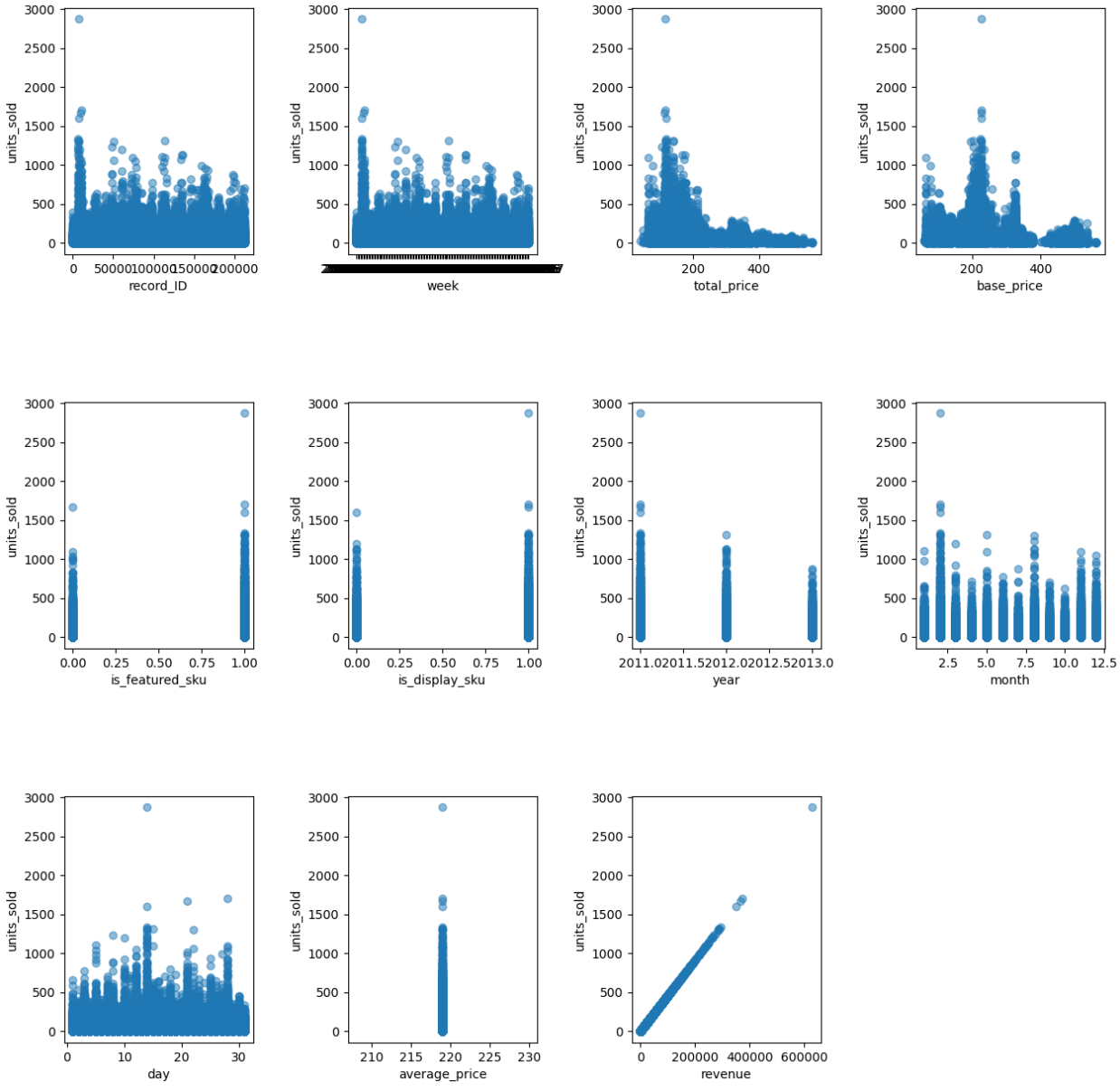
Exploratory Data Analysis

At first we performed **Principal Component Analysis (PCA)** to reduce the scope of the feature space and identify the principal components. Then, performed feature engineering which help us guide on how (or whether) to use the labels in the data. Then, we create **Feature correlation heatmap** to gain a high level view of relationships amongst the features. An interesting observations are as follows:



Another insight comes when we created a scatterplots of numeric features against units sold to identify patterns. Results shows:

- There's a strong positive correlation between units_sold and revenue (as expected).
- Columns such as year, month and day seems very useful to understand the seasonality and behavioral aspect.
- Columns such as is_featured_sku and is_display_sku appear quite similar and there's no difference as such.
- There are some outliers present in almost all columns.



Pre-processing and Training data

To further improve our data quality and make the data useful for machine modeling, we further explored the data by first drop insignificant columns, impute missing values, proper clean and transform the data in right format, allowing the machine learning model to learn effectively.

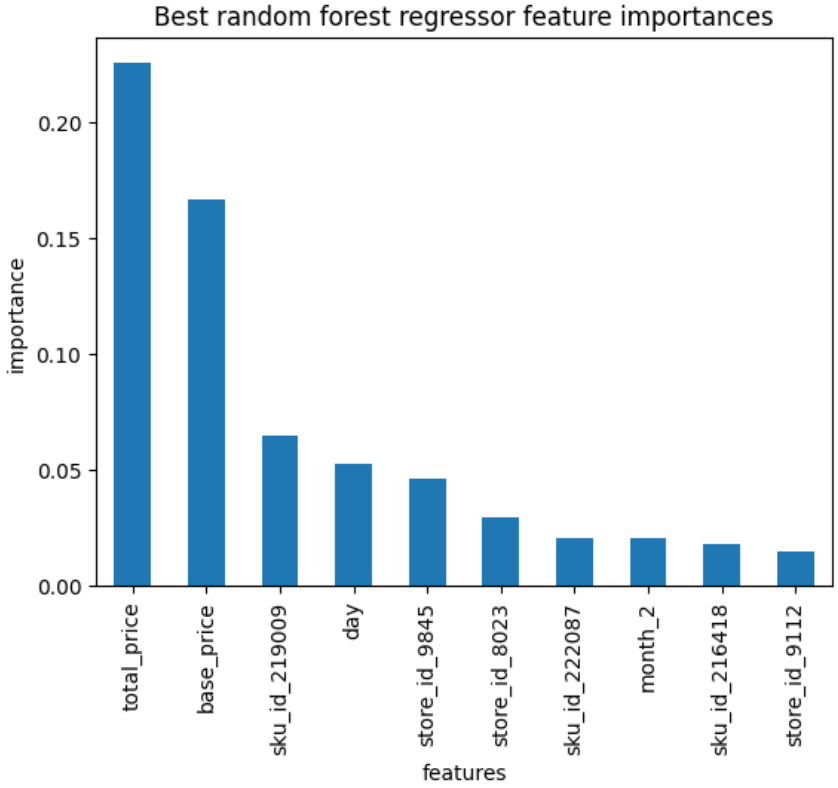
Drop insignificant columns such as 'week','is_featured_sku', 'is_display_sku', 'average_price', ‘revenue’.

Initiate feature engineering process by performing “One-Hot Encoding” on categorical columns such as "month", "store_id", "sku_id", “year".

Next step was to split the data into train and test dataset by partitioning the sizes with a 70/30 train/test split. This helps us learn the relationship between the input features and the target variable (in this case units_sold).

Modeling

In this step, we select and finalize the best model (Linear Regression model, Random forest regression model and XG Boost model) and calculate the mean absolute error using cross-validation. By using Random forest regression model, we are able to identify the dominant top features that are somewhat common with linear model.

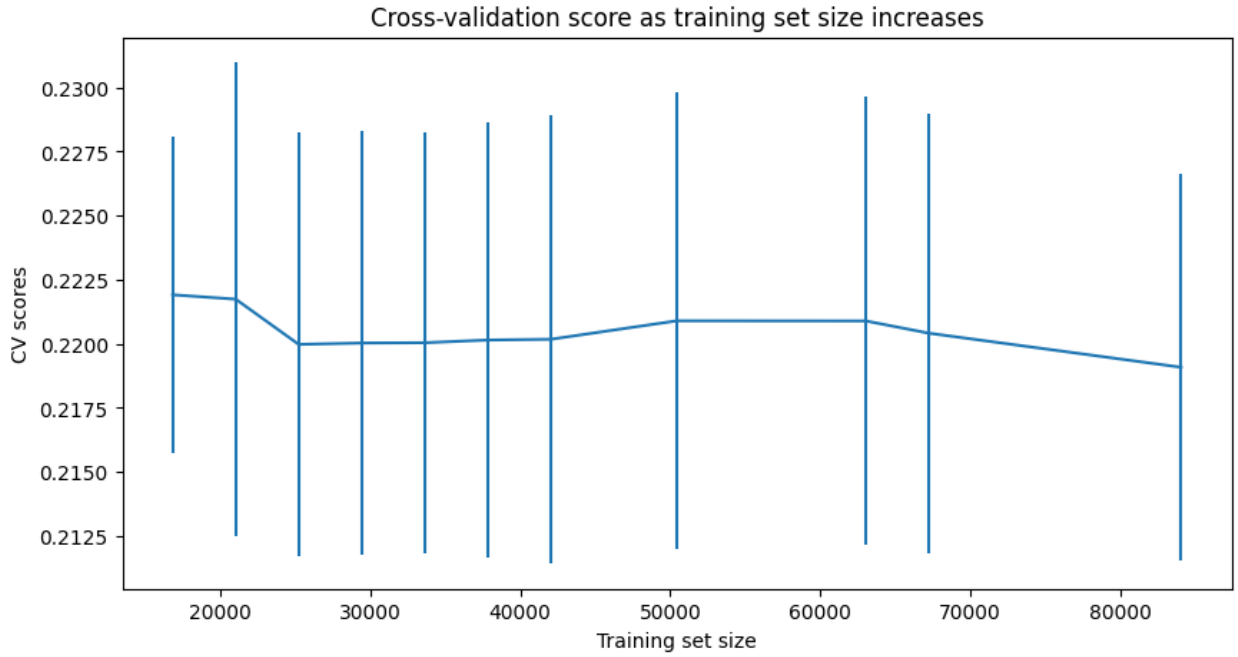


Features that came dominant in the modeling (random forest model) and affect demand the most are as follows:

- 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

The random forest model has a lowest cross-validation mean absolute error of 13.45 as compared to other models. It also exhibits less variability.

We also assess the quality of the data and it shows that there's an initial rapid improvement in model scores, but it's essentially leveled off by around a sample size of 25000 - 30000.This shows that the model scores are essentially leveled off throughout.



Recommendations

This Random Forest model can be further improved by:

- Adding more relevant features like external factors (economic indicators, holidays, etc.) to further improve the 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.
- Can include external data sources such as marketing campaigns, competitor actions to stay updated.
- Tuning the model periodically with fresh data to keep it up to date.