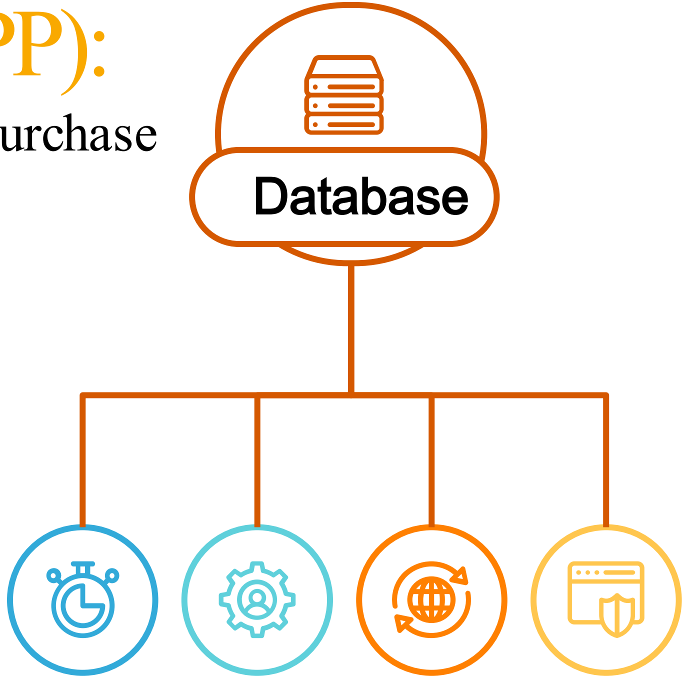
Product Price Predictor (PPP):

Tree-based Regression Model for Optimal XOX Purchase



Group 4 Members (in presenting order):

Fei Han

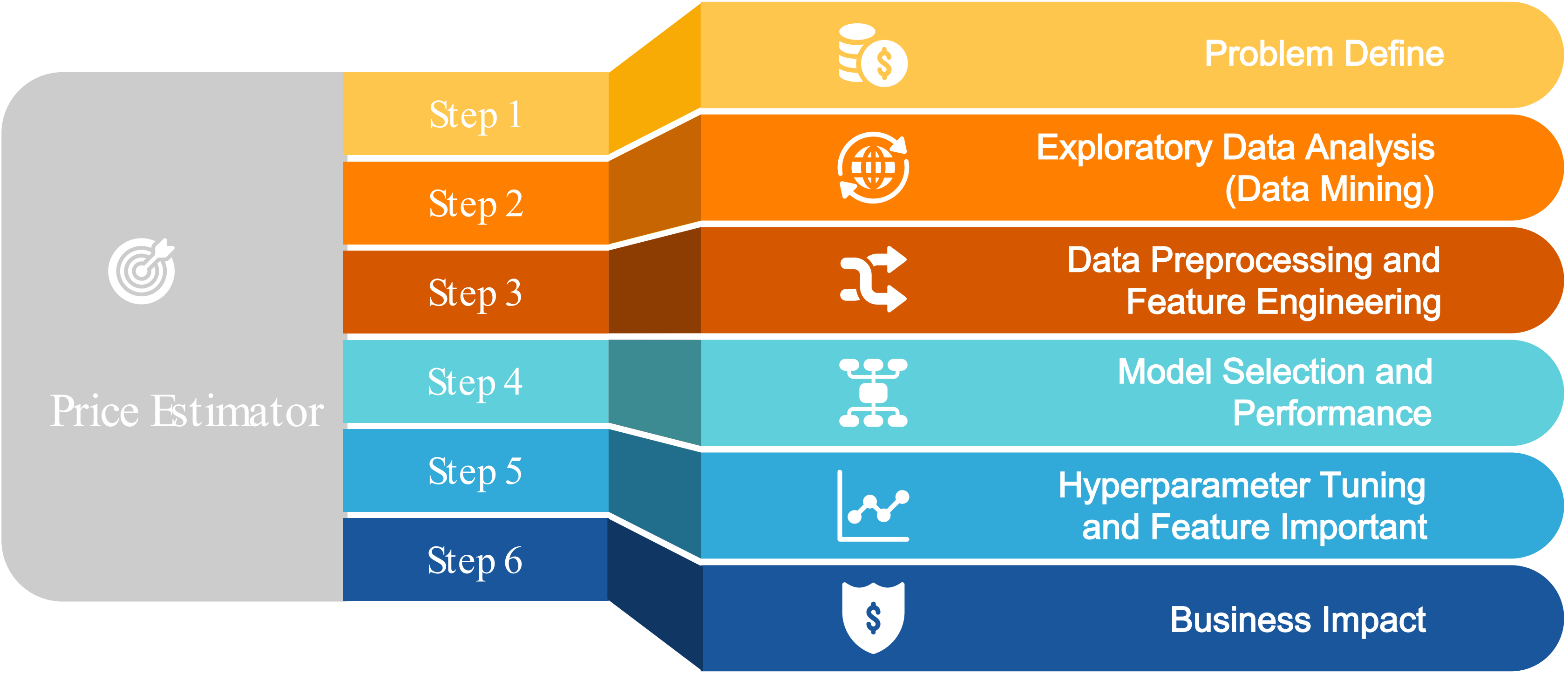
Chengkun Xing

Mingyue Zheng

Lili Chen

Mingjie Tang

How do we build the PPP model?



1. Problem Define

About us

We are an agency helping our customers purchase XoX from various makers.

Goal

To estimate the price of a XoX before we recommend it to our customers

Business service

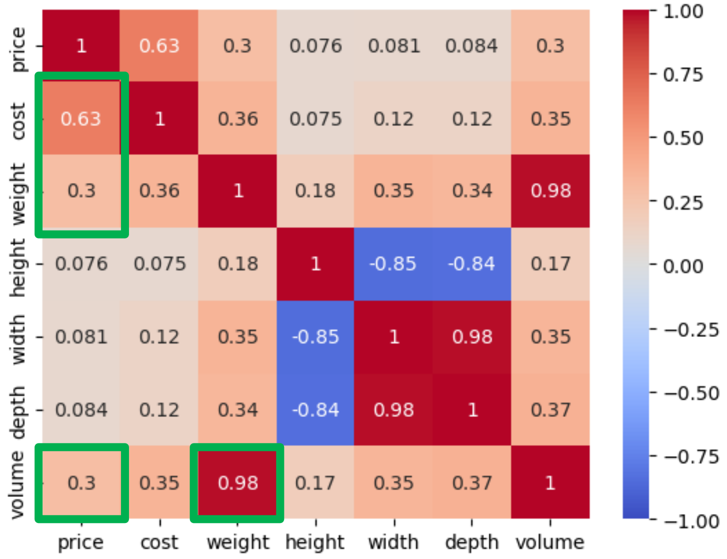
Provide business insights to explain the predicted price to our customers

Machine Learning service

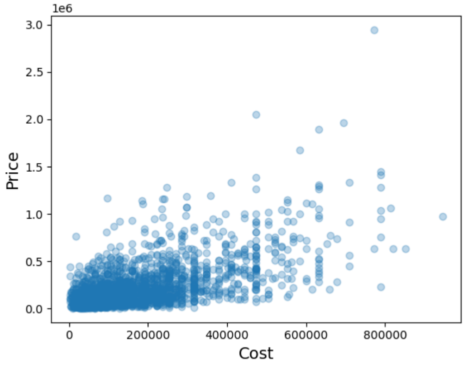
To build a machine learning model to accurately predict the price for a future purchase

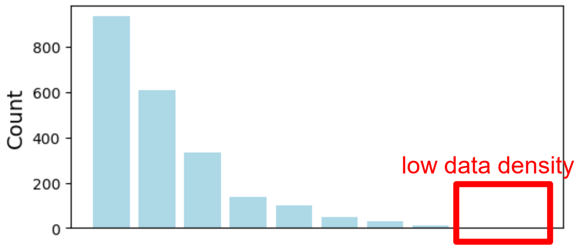
2. Exploratory Data Analysis (Data Mining)

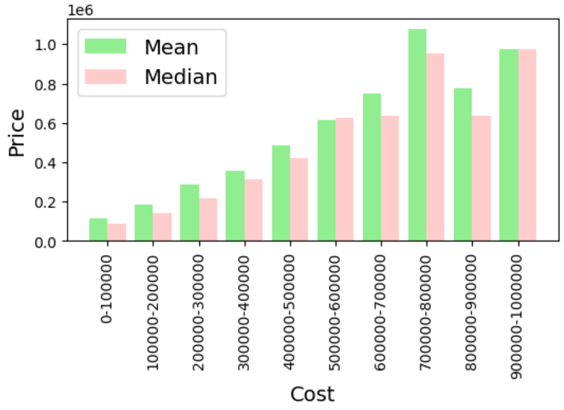
Correlations in numerical data



Analyze numerical data by statistical data binning



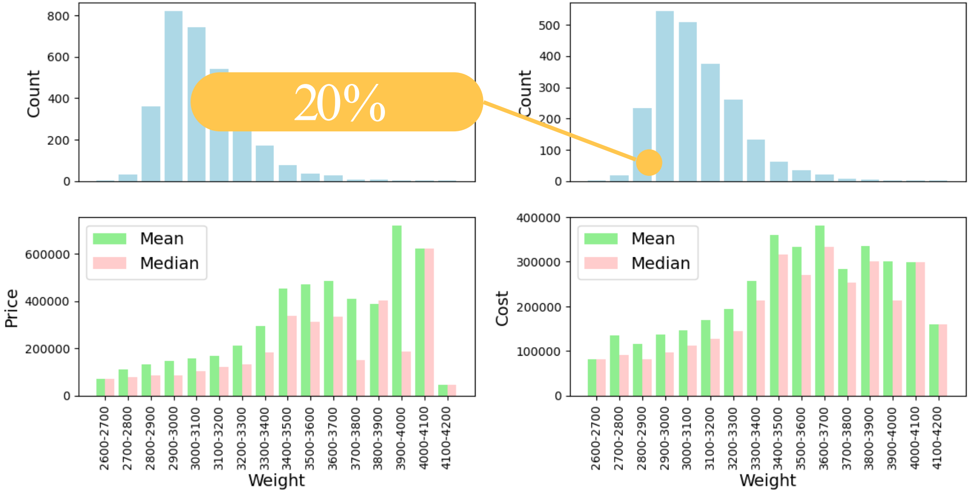


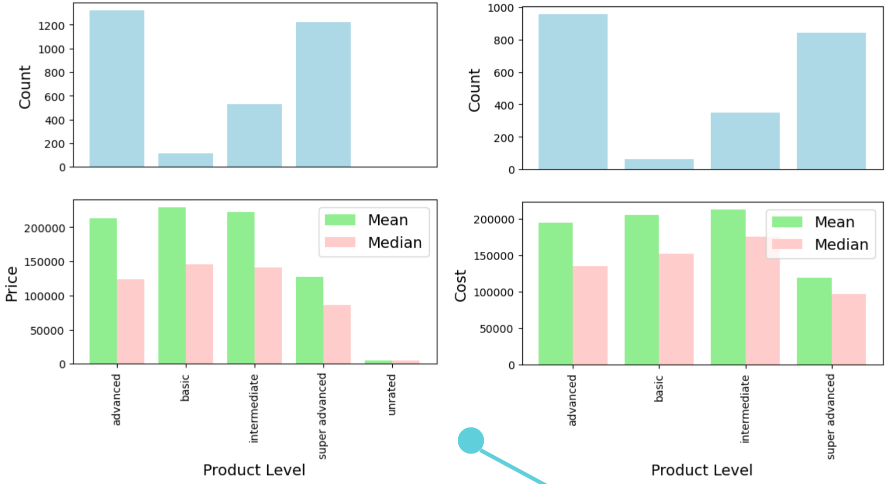


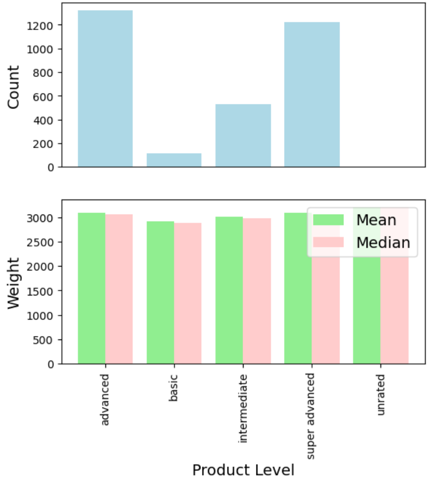
- Scatter plot: difficult to find the trend and pattern in data
- “Statistical data binning”: statistics along both x and y axes
- Capture the trend and pattern among high-data-density bins

Data Mining

- Relationship between price and weight is quasi-linear/ quadratic
- Cost and weight show a similar trend with price and weight

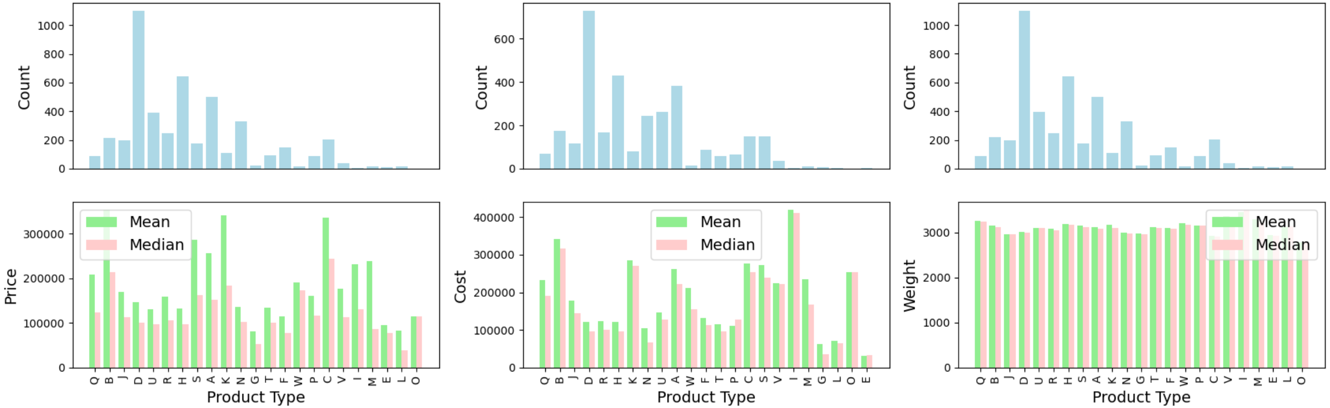






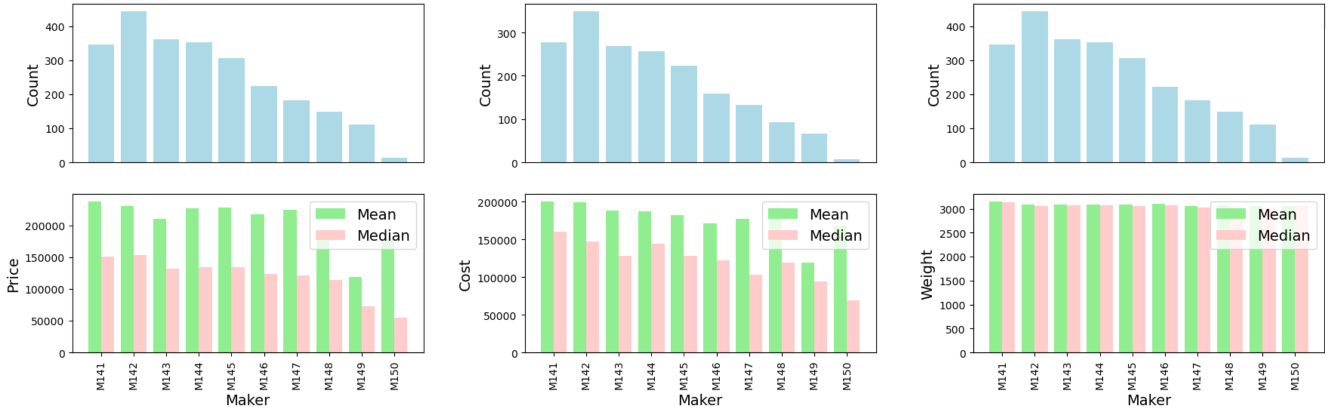
- Price and cost are relevant to the product level
- Weight is independent of the product level

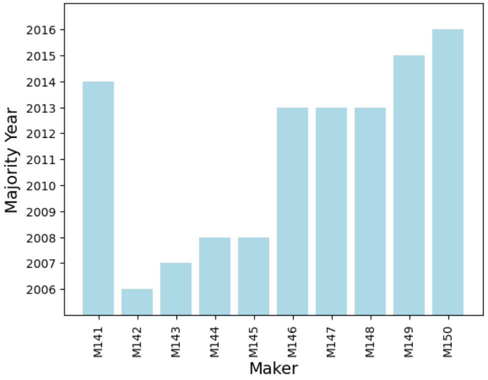
2. Data Mining



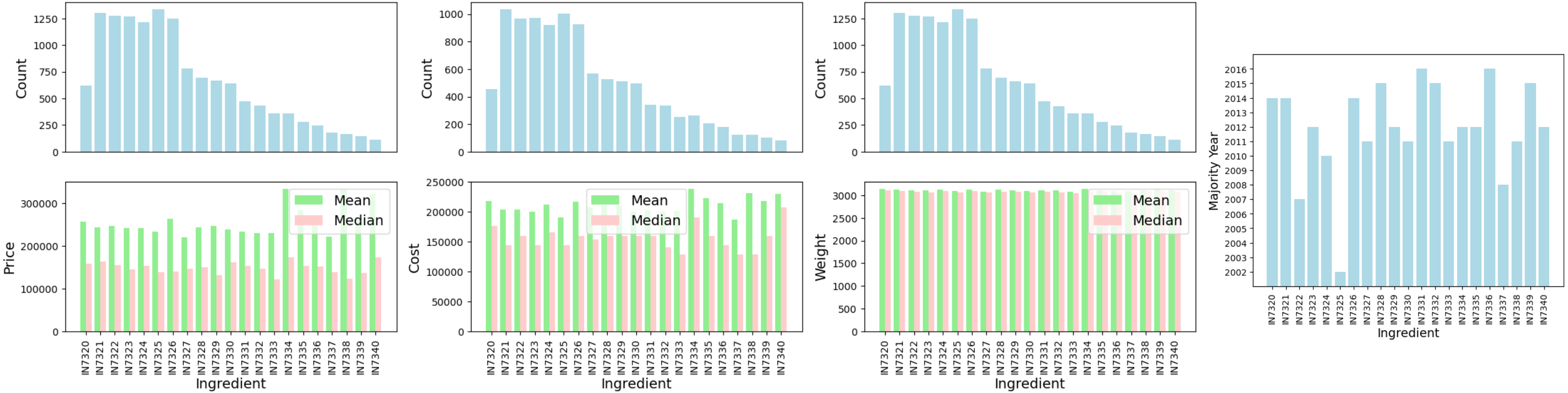
- Price and cost are relevant product type
- Weight is independent of product type

- Price and cost are somewhat relevant to maker
- Weight is independent of maker
- Ordinal numbers in maker data basically increase with year



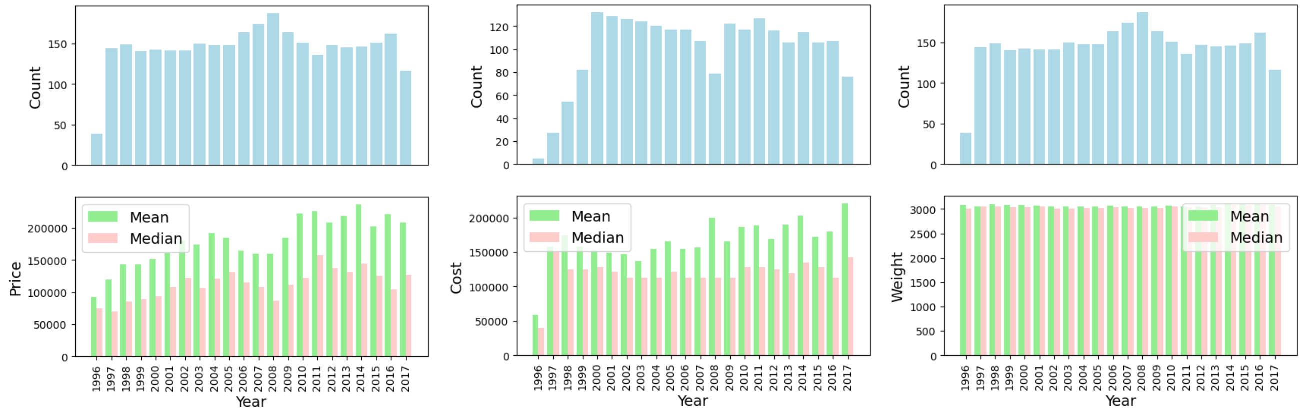


2. Data Mining

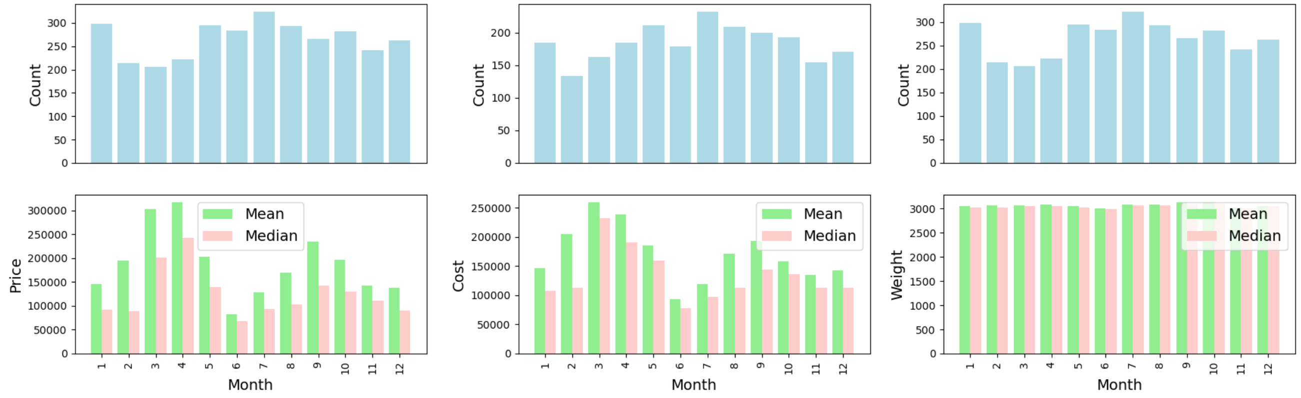


- Price and cost are weakly correlated with ingredient
- Weight is independent of ingredient
- Ordinal numbers in ingredient data have no correlation with year

2. Data Mining

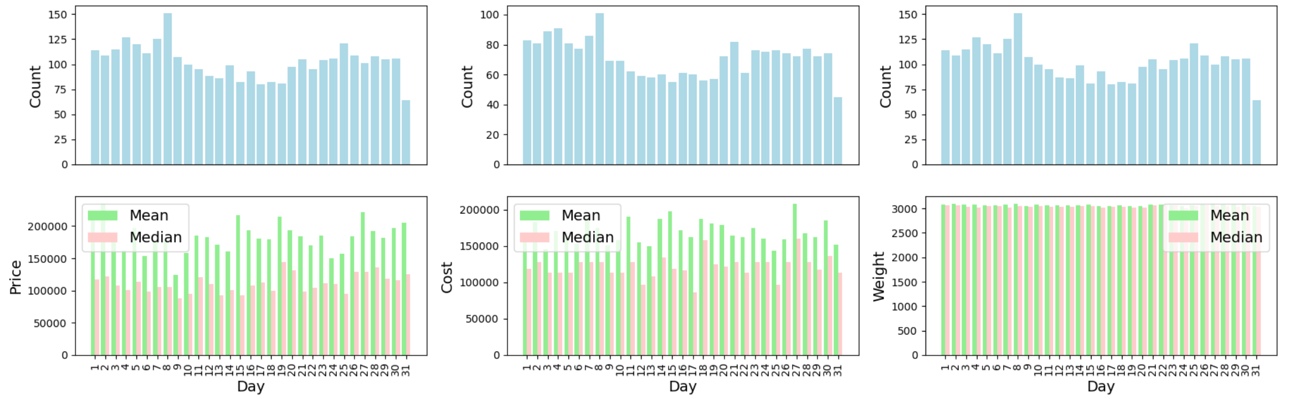


- Price shows a correlation with economic cycle (yearly inflation and economic crisis)
- Price of current year is correlated with those of past years
- Cost and weight are almost independent of purchase year

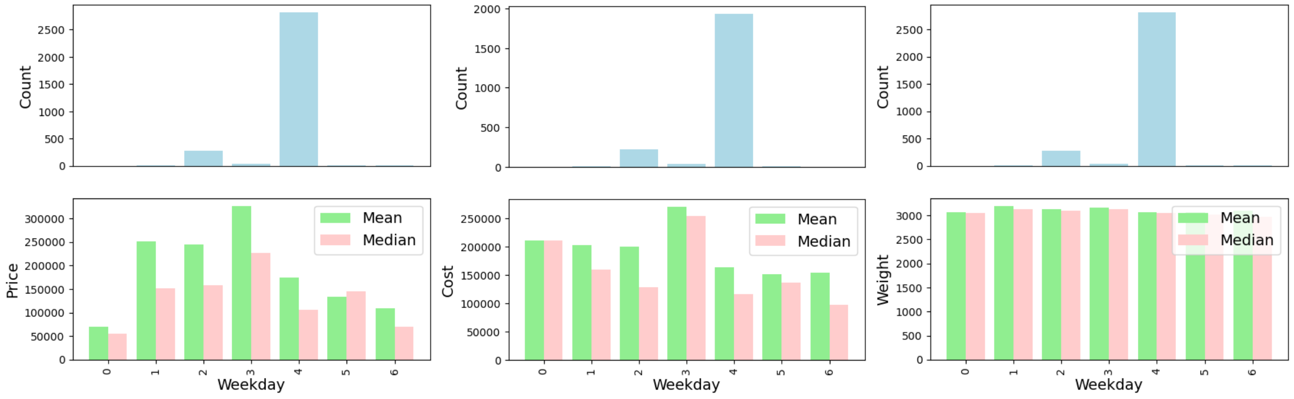


- Price and cost show a strong correlation with the season (high in spring and fall, low in summer and winter)
- Weight is independent of purchase month

2. Data Mining



- Price and cost are almost independent of purchase day
- Weight is independent of purchase day



- Most of transactions occurs on Thursday and Tuesdays
- On Thursday and Tuesdays, prices are different
- Weight is independent of purchase weekday

3. Data Preprocessing & Feature Engineering

0.3%
price

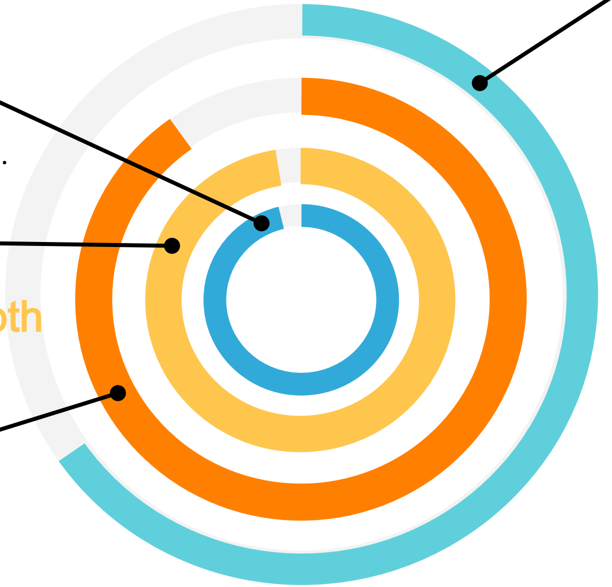
Drop the missing target.

0.1%
weight, width, depth

Drop a few lines.

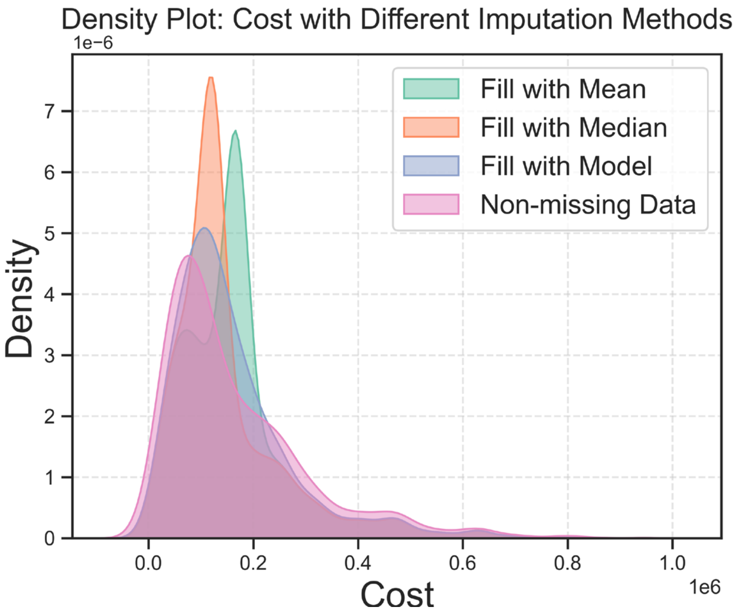
10%
ingredient

Label as “unknown”.



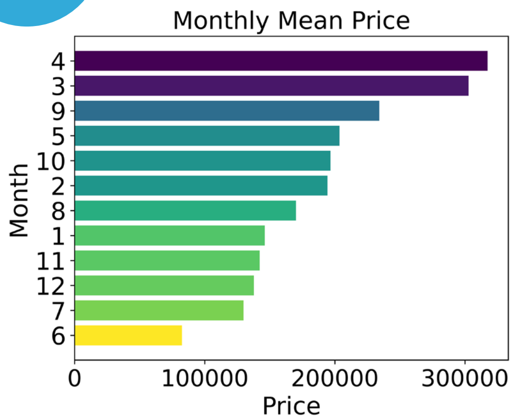
30%
cost

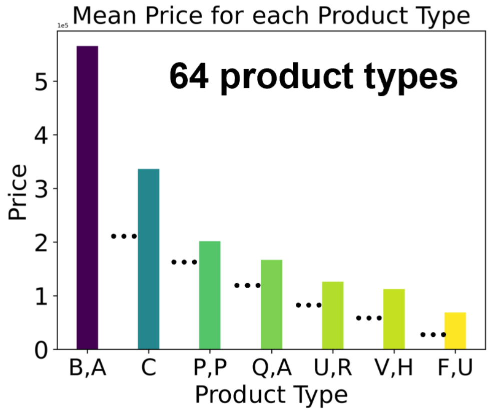
- Considerable missing fraction.
- Build a linear model to impute it based on other features.



3. Data Preprocessing & Feature Engineering

Categorical Features





Month

- Price varies a lot in different month.
- Directly used as categorical feature.

Product Type

Product level

- Price variability in Product Type and Levels.
- Number of categories is not too big.
- Create ***one-hot dummy*** features.

3. Data Preprocessing & Feature Engineering

Numerical Features

Get rid of collinearity

01 High collinearity between depth, width, height.

02 Combine height, width and depth into a single feature: volume.

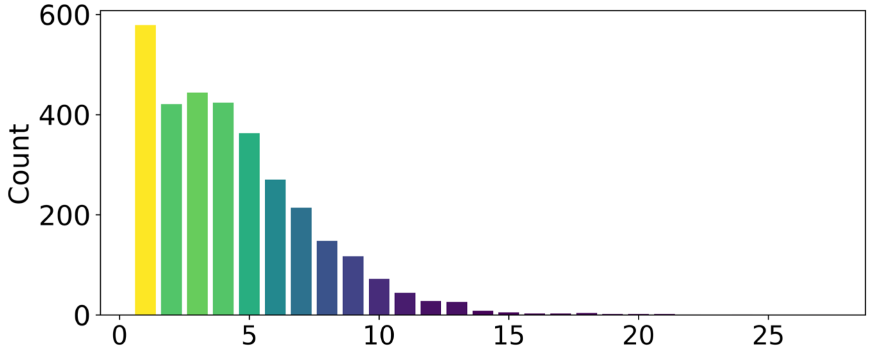
03 High collinearity between weight and volume.

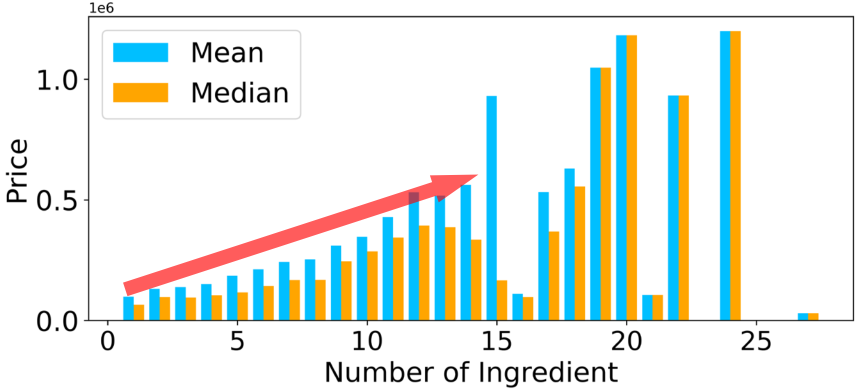
04 Drop volume and just keep weight.

3. Data Preprocessing & Feature Engineering

Numerical Features

eg: IN732054,IN732059
2 Ingredient Number



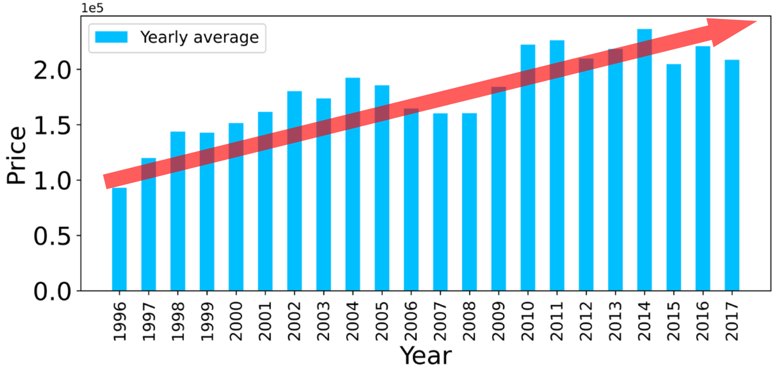


Ingredient Number

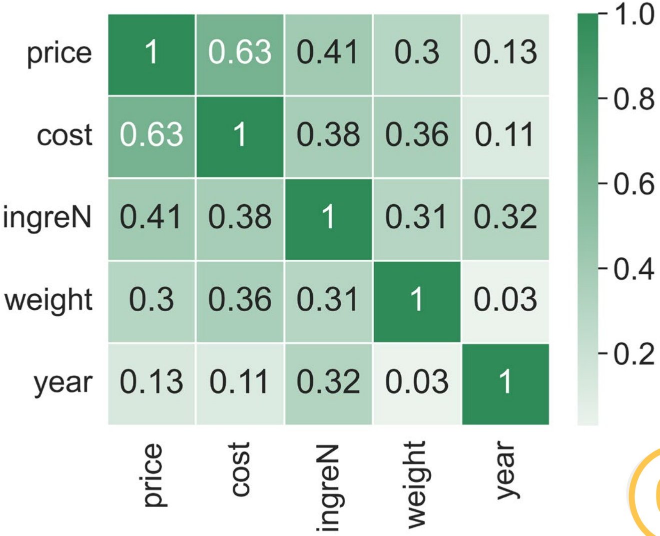
- More numbers of ingredients, higher price.
- Price drops when the number of ingredients is larger than 13, but the count is also very small.

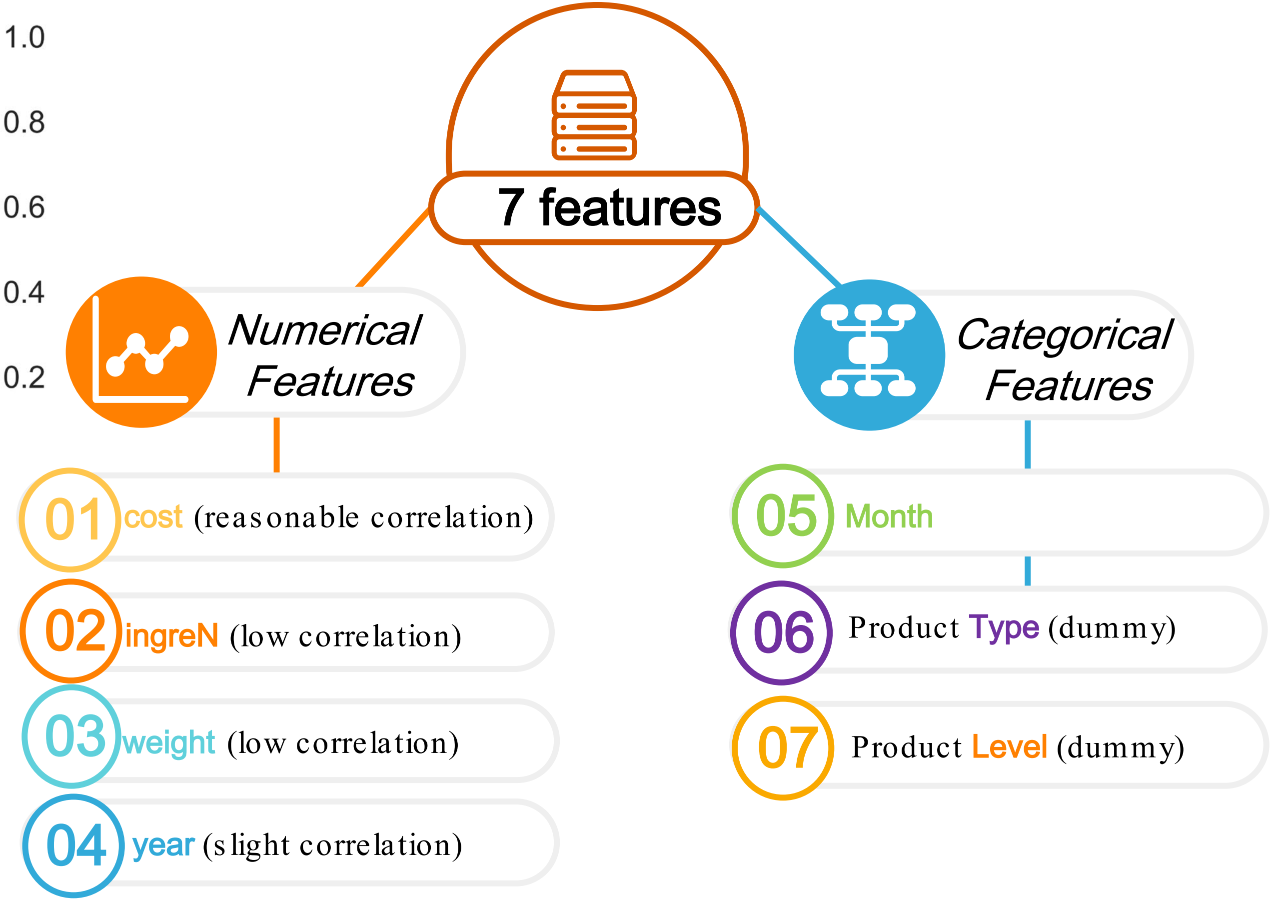
Year

- Price increases a little with the year.
- May not work in Tree model.



3. Data Preprocessing & Feature Engineering

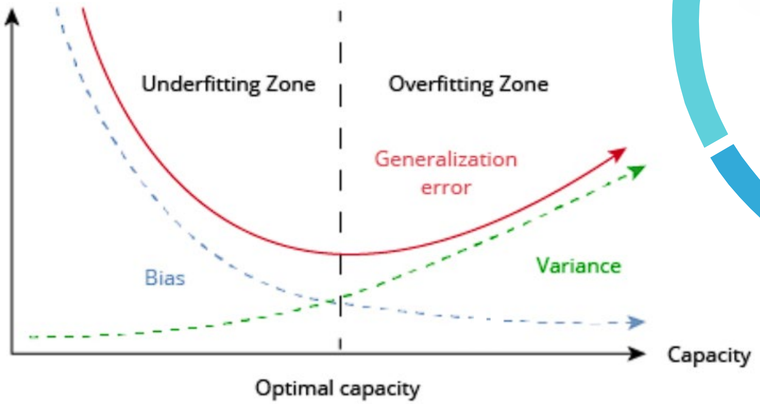


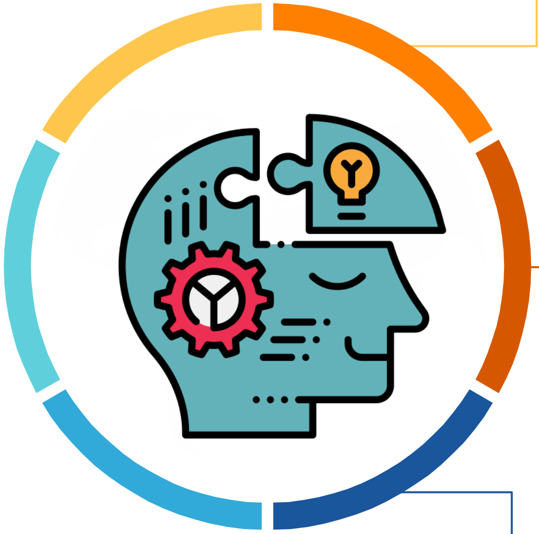


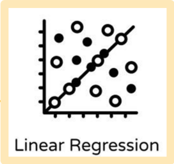
4. Model Selection and Performance

Why did we choose tree-based models?

- Handle non-linear relationships effectively
- Robustness to outliers and missing data
- Feature importance
- Ensemble learning
- Scalability and parallelization





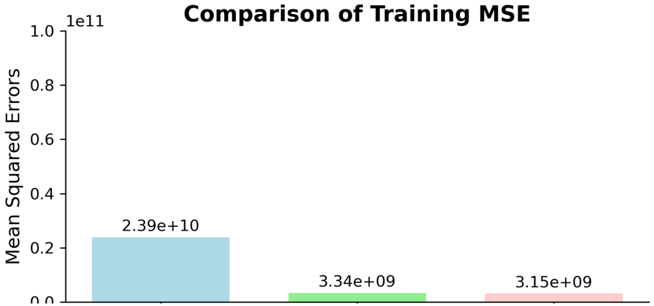


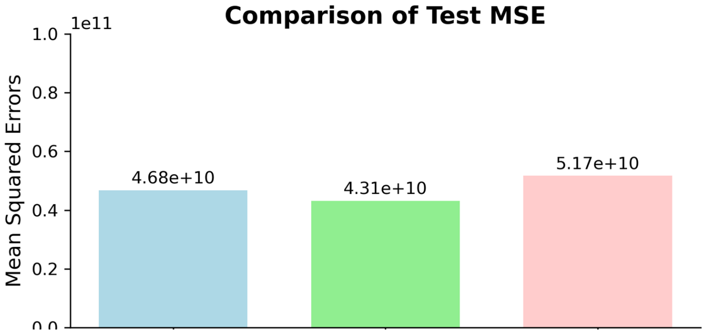
Linear Regression
High bias
Low variance

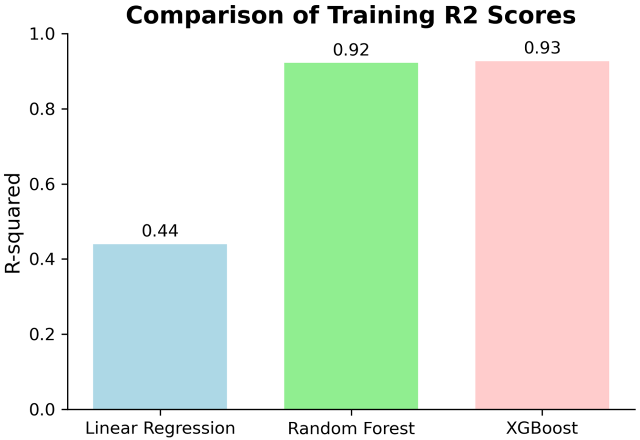
Random Forest
Low bias
Low variance

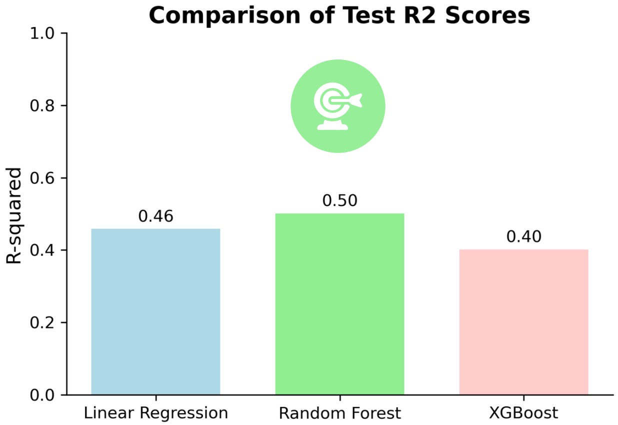
XGBoost
Low bias
Low variance

4. Model Selection and Performance









5. Hyperparameter Tuning & Feature Importanc

Cross_Validation Method: GridSearchCV

Hyperparameters tuning results for Random Forest: (r2_score)

max_depth =11,

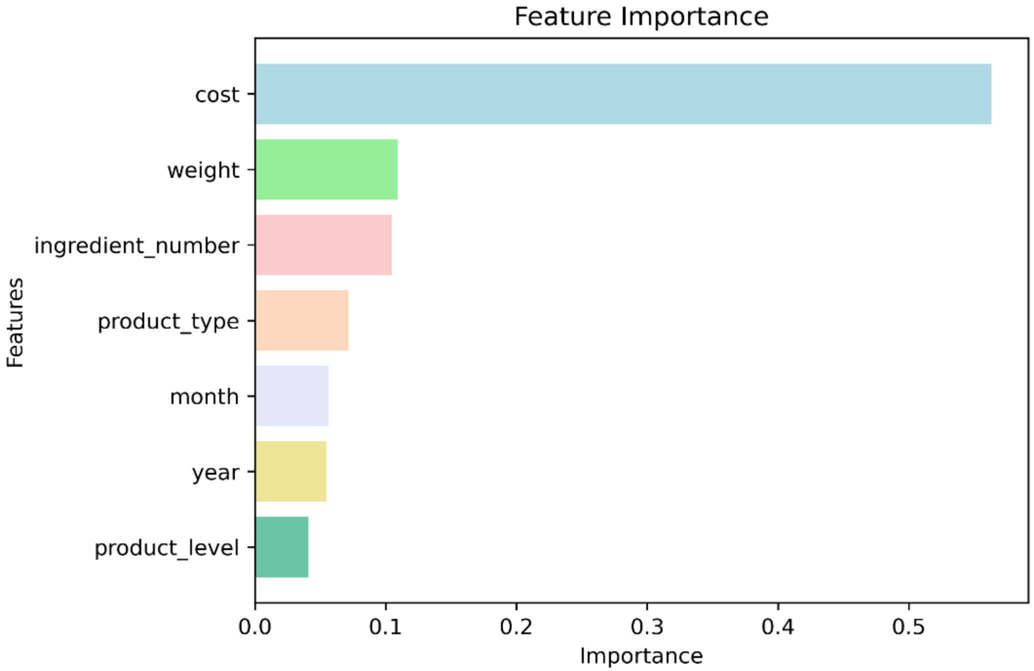
min_samples_leaf=3,

min_samples_split=2,

n_estimators=500

7 features

5.Hyperparameter Tuning & Feature Importance



- Random Forest (R2=0.52) vs. linear regression (R2=0.46) benchmark
- Key features: Cost, Weight, Ingredient Number, Product Type, Month, Year
- Higher cost, weight and ingredient number -> higher price
- Cheaper prices in June, July, Dec., Nov., and Jan.

5. Business Impact

- **What is the Benchmark?**

- Linear regression model with cost as only input
- Linear model tends to have high bias and it is suitable candidate for Benchmark.

- **How to qualify the ML model improvement?**

$$\textit{Residual Percentage} = \frac{True\ Price - Predicted\ Price}{True\ Price}$$

Measures how far off are the price predictions

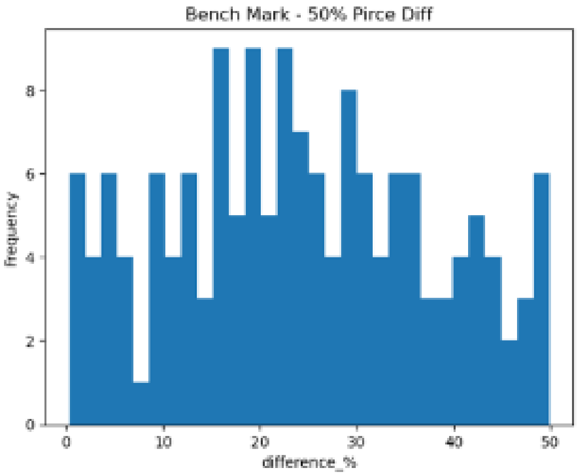
5. Business Impact

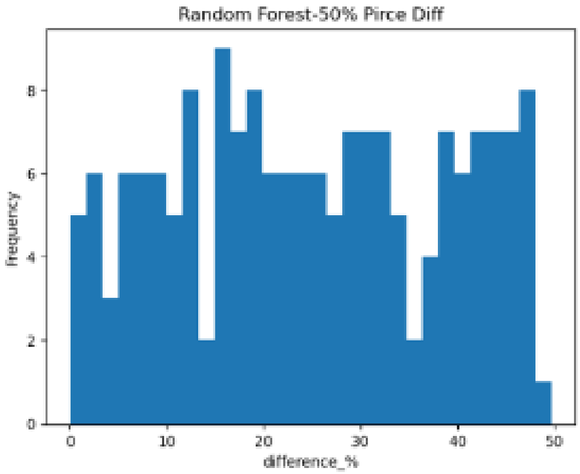
Comparison between Benchmark and Tree Model

50% Residual Percentage

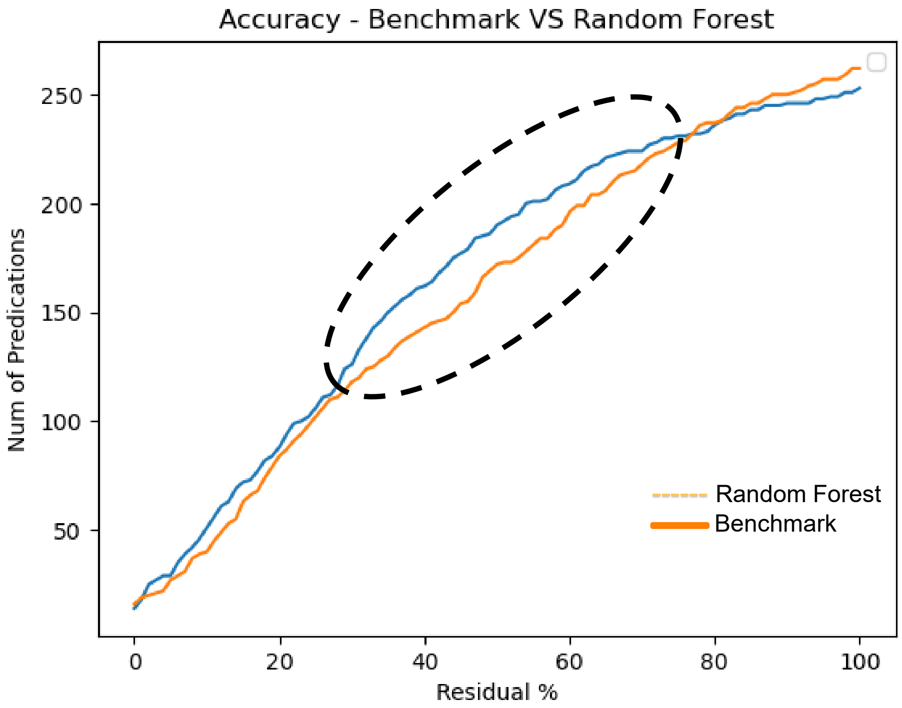
- Benchmark - 154 accurate predictions
- Random Forest - 175 accurate predictions
- Random Forest Produces **21** more accurate predictions **(Performance Index – PI)**

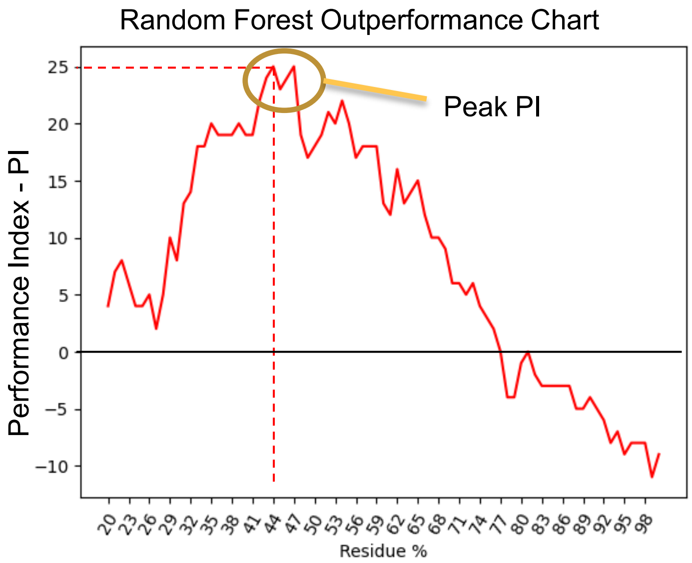
13% More Accurate Predications





Quantify Model Performance

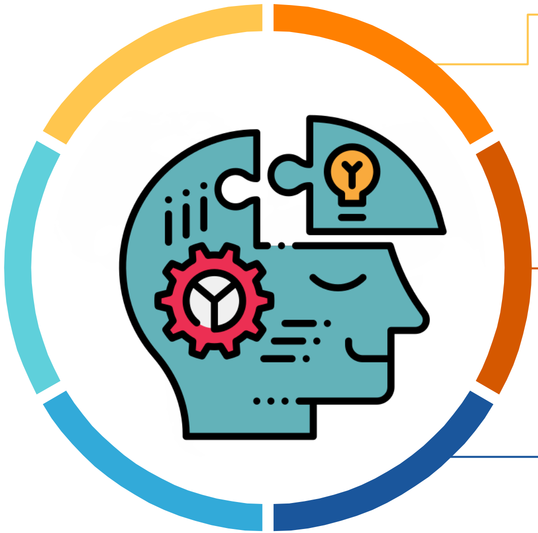




- Random Forest Model has its peak Performance Index (PI) when residual % is 44%.
- Tree model produces **25 more accurate price predictions.**
- Approximately **18% increase from Benchmark.**

5. Business Impact

Recommendations



Improve data entry quality-Cost has 30% missing value

Missing Benchmark definition

Missing definition of “Accurate Predication” - Need to further communicate with customers to better understand and address their needs and concerns.