Walmart

Purpose of Analysis: What are the deterministic factors that influence Weekly Sales? Feature Importance: How each feature contributes to the predictive performance of machine learning models in terms of impurity like variance.

Variables:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Store | Weekly Sales | Holiday Flag | Temperature | Fuel  Price | CPI | Unemployment  (%) |
| Type | Categorical | Numerical | Categorical | Numerical | Numerical | Numerical | Numerical |
| Min | 1 | 2.099862e+05 | 0 | -2.06 | 2.472 | 126.064 | 3.879 |
| Max | 45 | 3.818686e+06 | 1 | 100.14 | 4.468 | 227.232807 | 14.313 |
| Mean | \ | 1.046965e+06 | \ | 60.663782 | 3.358607 | 171.578394 | 7.999151 |
| Std | \ | 5.643666e+05 | \ | 18.444933 | 0.459020 | 39.356712 | 1.875885 |
| Med | \ | 9.607460e+05 | \ | 62.67 | 3.445 | 182.616521 | 7.874 |
|  | Date |
| Type | Datetime |
| Range  (Y/M/D) | 05/2010-10/2012 |

Question: How to convert Store and Date into variables?

Assumptions:

1. Location of Store will influence Weekly Sales.

Method: groupby .agg(mean) to see if there are huge difference across different store

Premise: They all share the same/similar data size

1. Different Seasons will influence Weekly Sales.

Method: Create Categorize variables based on date

Variable Distribution:

A screenshot of a graph

Description automatically generated

Skewness are found in several variables(Weekly\_sales,Unemployment, CPI)

QQ plot for Numerical Variables:

A graph of different values

Description automatically generated

Only temperature and Fuel Price close to normal distribution. That makes sense(season). Other variables are generally not normally distributed.

Models Choice:

1. Linear Regression:

When reject Linear regression model:

1. Non-normality and skewness

Action: Try to normalize data

Method: Minmax Normalization, Log-normalization(Weekly\_sales), boxcox. And then detect the normality by using Shapiro Test

1. Non-Normality of residuals or Homoscedasticity (The variance of residuals across all levels of independence variables)
2. Multicollinearity
3. Non-linear Models Selection:
4. Polynomial: Relationship between features and target value seems to be curved or quadratic.

Cons: Overfitting

**Steps to evaluate:**

1. **Scatter plot to show relationship**
2. **Fit polynomial trend line**
3. **Fit Polynomial regression and evaluate performance**
4. Random Forest: High Feature Interaction

Principle: Bootstrap data( with all the features) each time; For each bootstrap example, a decision tree is built. For each node, the feature selection is randomly.

Problem for create dummies:

1. Feature importance is distributed( 85% of variables are dummies created by one variable) , which means they are correlated with each other and random forest will not penalize for correlation because it just ensemble the trees, weight all equally.
2. Random forest tends to select these dummies more frequently.

Cons: Less interpretable

1. Gradient Boosting(LightGBM,XGboost): Learn from the previous trees to correct earlier errors, more weight for later trees.

Cons: Computation intensive

Based on our data: Highly and Mildly un-normalized data.

First, draw correlation to see if there are high-correlated variables influence the performance of Random Forest. (Using normalize data to easily compare the feature importance)

A screenshot of a graph

Description automatically generated

Not Highly correlated variables. Then Put all variables into Random Forest and get the feature Importance.

A screen shot of a computer

Description automatically generated

It seems that the seasons and holidays are much less important than Numerical Values. (If we choose top 5 descending).Therefore, try to combine the categorical variables to step forward on our assumptions. ( Season holiday influence weekly sales? )

Create Feature interactions. Draw correlation heatmap and make VIF test( avoid redundancy)

A black screen with white text

Description automatically generated

A screen shot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

Then run random forest again: Use shap to show the detailed breakdown of contribution.

A graph with different colored lines

Description automatically generated

A screen shot of a computer

Description automatically generated

Just to be accurate, use XGboost(grid search) to see if they return the similar feature importance.( More accurate)

A screenshot of a graph

Description automatically generated

Therefore, since they both return similar feature importance, it means that these top features have a significant and consistent relationship with the target variable.

The next step is:

Use cross validation(cv=5) and feature combinations across the top five features to improve the model fit( MSE and R^2) based on certain features.

A black background with white text

Description automatically generated

Unemployment is the dominant indicator for weekly sales, much more stranger than other, higher predictive than others

What else we can do?