Feature selection methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable. What algorithms can be used to automatically select the most important features (regression, etc..)? Describe at least 3?

Answer:

1. Pearson Correlation: This is a filter-based feature selection method. We check the absolute value of the Pearson’s correlation between the target and numerical features in our dataset. And keep the top n features based on this criterion.
2. Chi-Squared: We calculate the chi-square metric between the target and the numerical variable and only select the variable with the maximum chi-squared values.
3. Recursive Feature Elimination: select features by recursively considering smaller and smaller sets of features.

Explain data leakage and overfitting (define each)?  
Explain the effect of data leakage and overfitting on the performance of an ML model.

Answer:

Overfitting occurs when the model performs too well on the training data but poorly on the new data points while the goal is to maximize its accuracy on the unseen data points. This means that the model fails to generalize to unseen data (i.e., the trained model is not able to adapt to new data).

Data leakage occurs when an information is leaked/introduced in the training dataset from a data point that would not be expected to be available during the training phase (i.e., having test dataset included accidently in training dataset).

When data leakage occurs, the model performs too well on test data set. Data points are leaked from test set into training set, which means the model will predict something it saw (which is expected to be unseen data points).

Explain what our outliers in your data?  
Explain at least two methods to deal/treat outliers in your data?

Answer:

Weekly\_Sales, MarkDown2, 3, 4, 5 has outliers in our data. Take Weekly\_Sales as an example, IQR(Median) = 10357.32, Q1 = 3502.09, Q3 = 31647.36. Maximum = Q3 + 1.5 IQR = 31647.36+1.5\*10357.32 = 47183.34. From .description(), we can see there exists data points’ Weekly\_Sales larger than 47183.34, which is the outlier.

To deal with outliers, we can cap outliers at a certain value above the 90th percentile value or floored at a factor below the 10th percentile value. Median imputation is another option.

What is feature scaling and why is it important to our model?  
Explain the different between Normalization and Standardization?

Answer:

Feature scaling is a method used to normalize the range of independent variables or features of data. Most of the time, the dataset will contain features highly varying in magnitudes. If we apply machine learning algorithm on dataset without scaling, the big scaled feature becomes dominating which might be incorrect. To suppress this effect, we need to bring all features to the same level of magnitudes. This can be achieved by scaling.

Normalization means rescales the values into a range of [0,1]. Standardization means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).