Feature Generation

Feature generation is also known as feature construction, feature extraction or feature engineering. While these differ in definition a little, we are going to focus on this one - Feature generation is the process of creating new features from one or multiple existing features, generally for use in statistical analysis.

Feature generation helps us achieve :

1. Dimensionality reduction
2. Accuracy improvement

When dimensionality reduction is the main goal, the resulting feature space has less features than the original one. When accuracy improvement is the main goal, the resulting feature space most likely has more features than the original one.

We are mainly interested in improving the accuracy of the model. Dimensionality reduction is not our priority as feature generation is an input to feature selection, which takes care of it. However, it is important to ensure that we do not generate a huge amount of new features.

Feature generation also helps us improve the performance when there is a feature interaction, that is, when two or more features are not relevant or correlated to the dependent feature, but together have an influence on the dependent feature. For instance, given a dataset of housing prices, combining the length (in ft) *l* and width (in ft) *w* of the plot, which do not have an influence on the price of the house (dependent feature) directly, by *sqft = l*w* would result in a new feature, which gives us square footage of the plot, which plays a significant influence on the price.

Feature Selection

Once you have generated new features in addition to the existing features, you are left with a relatively wide set of features. Choosing the right subset of features from this feature set is essential to optimize your model performance, that is, improve or maintain model accuracy and simplify its complexity.

The number of possible subsets of a set of n features is 2^n. Testing the model for each of these 2^n subsets is not possible, so we must resort to a better method of feature selection.

Feature selection techniques that are effective are -

- Univariate selection
- Feature importance
- Correlation matrix using heatmap

Univariate Selection

Statistical methods are used to select the features that have the highest relation with the dependent variable/ output variable.

The scikit-learn library provides the SelectKBest class that can be used with a suite of different statistical tests to select a specific number of features.

In the example below, we have used the chi-squared (χ^2) statistical test to select 10 of the best features from the Mobile Price Range Prediction Dataset. Pearson's chi-squared test is a statistical test that measures how expectations compare to model results.
https://www.kaggle.com/iabhishekofficial/mobile-price-classification?select=train.csv
Independent variables in this dataset are:
battery_power: Total energy a battery can store in one time measured in mAh

blue: Has Bluetooth or not

clock_speed: the speed at which microprocessor executes instructions

dual_sim: Has dual sim support or not

fc: Front Camera megapixels

four_g: Has 4G or not

int_memory: Internal Memory in Gigabytes

m_dep: Mobile Depth in cm

mobile_wt: Weight of mobile phone

n_cores: Number of cores of the processor

pc: Primary Camera megapixels

px_height

Pixel Resolution Height

px_width: Pixel Resolution Width

ram: Random Access Memory in MegaBytes

sc_h: Screen Height of mobile in cm

sc_w: Screen Width of mobile in cm

talk_time: the longest time that a single battery charge will last when you are

three_g: Has 3G or not

touch_screen: Has touch screen or not

wifi: Has wifi or not

Dependent variable is :

price_range: This is the target variable with a value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

Feature Importance

In this technique, we use the feature importance property of the model. Feature importance gives you a score for each feature of your data, higher the score more important or relevant is the feature towards your output variable.

Feature importance is an inbuilt class present in Tree Based Classifiers, we will be using Extra Tree Classifier for extracting the top 10 features for the dataset.

(Extra Tree Classifier gives more diverse trees than Random Forest Classifier. It also works better on noisy features than)

Correlation Matrix with Heatmap

Correlation states how the features are related to each other or the dependent variable (in this example, it is price range).

Correlation is positive when an increase in the value of a feature results in an increase in the value of the dependent variable) and it is negative when an increase in the value of a feature results in a decrease in the value of the dependent variable).

Heatmap makes visualizing which features are most related to the dependent variable easier.

Task

Plot the correlation matrix using the heatmap for the given dataset.