SPRINGBOARD - DATA SCIENCE CAREER TRACK

REPORT ON DATA WRANGLING

HOUSING PRICES-ADVANCED REGRESSION TECHNIQUES

INTRODUCTION

This Capstone Project -1 is on Housing Prices: Advanced Regression Techniques. This dataset has been picked up from Kaggle. We would be cleaning , analysing and visualising the data using Python. Finally we will see the different Models that have been developed to predict Housing Prices in the best possible manner.

PROBLEM STATEMENT

Prediction of sale prices.

The problem for the buyer is knowing the exact amount for the purchase price of the house. It becomes crucial to know the levers that drive the price and develop a model to predict them with best accuracy.

DATA ACQUISITION

The Ames Housing dataset part of Kaggle competitions is used. The dataset has 79 explanatory variables describing every aspect of residential homes in Ames, Iowa.

DATA WRANGLING

Step1- Classifying dtypes

There are 1460 observations. There are 81 variables of int,float and object dtypes.

We classify data types into 2 classes i.e numeric and categorical. We also need to identify ordinal variables and decide which class we can place them in. Following lists are created:

NUMERIC (‘numcollist’)

CATEGORICAL(‘categlist’)

Step- 2 Imputation of Nan and Zero values

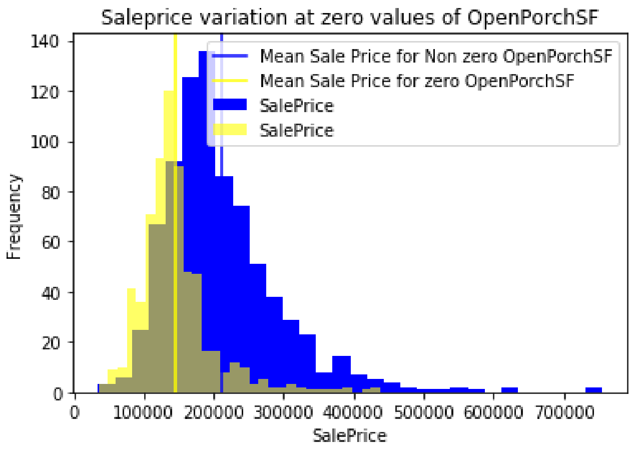
Nan values in Categorical variables:

For the Nan values we create a separate category 'Absent'.This would help to capture any new information such as average Sale Price in case a feature is absent. Variables like PoolQC, Fence, MiscFeature and Alley have a very high Nan. These may not be useful at all. However, we cannot make similar assumptions about population data. So we retain them for whatever little information they may provide.

Nan values and Zeros in Numeric variables:

More than the problem of Nan, we have a problem of zero values, as they were more in numbers. This could mean that a feature is not present. We tried to see the impact of a numerical feature being absent, on the 'SalePrice' variable by plotting the histogram of 'Sale Price'.

For example: In the histogram we found that Average SalePrice is much lower where OpenPorchSF is not present.



We treat Nan's and zero's as follows:

1. We create a seperate variable for each numerical column and label the absence of feature i.e 0 value as 1, and presence of a feature as 0.
2. We replace zeros with Nan and impute Nan's by linear interpolation.

Step-3 Identifying Outliers

1. We observed many values above z score of 3 in the target variable i.e Sales Price.
2. There were 22 such observations.
3. Before making a decision to remove or transform these outliers we looked at the linear relationship between Sale Price and GrLivArea.

4. Out of these 22 values most of the values are following a linear trend with GrLiveArea.But these values are high. If we normalize the data, this could be taken care of. However 2 points at bottom right are not following the trend. So these 2 points are outliers.
5. We later applied graphical tools to identify outliers such as residual plot