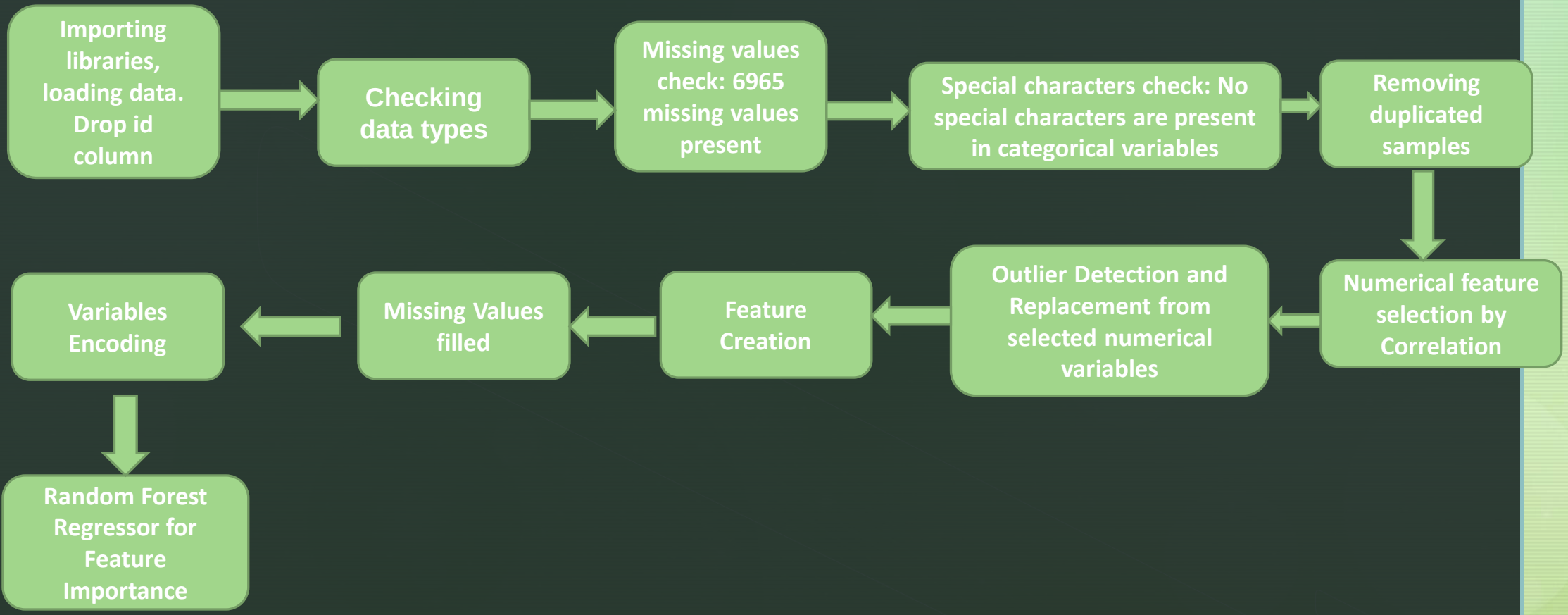
Assignment For: Home.LLC

Important Features for predicting House Prices in USA

Dataset

- Total no. of variables: 80
- Numerical variables: 37
- Categorical variables: 43
- Dependent variable: SalePrice
- No. of samples: 1460

Procedure


Importing libraries, loading data. Drop id column
Checking data types
Missing values check: 6965 missing values present
Special characters check: No special characters are present in categorical variables
Removing duplicated samples
Numerical feature selection by Correlation
Outlier Detection and Replacement from selected numerical variables
Feature Creation
Missing Values filled
Variables Encoding
Random Forest Regressor for Feature Importance

Numerical Feature Selection
Heatmap of Numerical variables

- Criteria for selection:
- Correlation with target variable SalePrice > 0.5
- Selected variables:

OverallQual,

GrLivArea,

GarageCars,

GarageArea,

TotalBsmtSF,

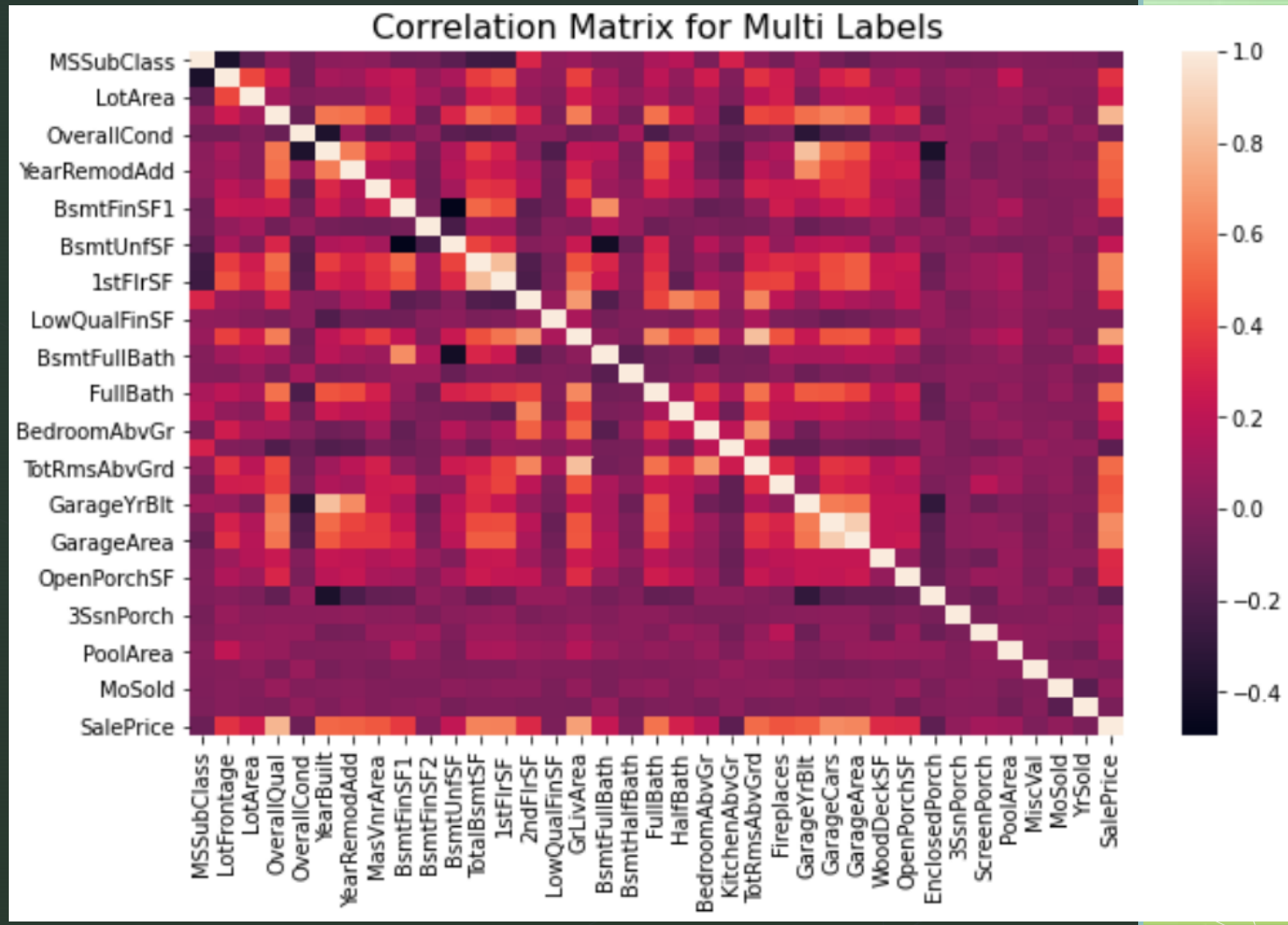
1stFlrSF,

FullBath,

TotRmsAbvGrd,

YearBuilt,

YearRemodAdd



Outlier Detection of Selected variables

- No outliers are present in variables: FullBath and TearRemodAdd. Outliers are present in variables OverallQual, GrLivArea, YearBuilt,TotRmsAbvGrnd, 1stFlrSF,TotalBsmtSF, GarageArea, GarageCars.
- Outliers replaced by median values as data is skewed and median is not affected by outliers whereas mean is affected by outliers.
- Inter quartile Method: Cut-off at 25% and 75% quantile and lower = q25 - 1.5*iqr; Upper = q75 + 1.5*iqr
- Outliers outside lower and upper bounds are replaced by median

Feature Creation

- TotalSF' = 'TotalBsmtSF' + '1stFlrSF' + '2ndFlrSF
- TotalNumberOfRooms' = 'BedroomAbvGr' + 'KitchenAbvGr’ + 'TotRmsAbvGrd
- TotalNumberOfBathrooms' = 'FullBath' + 0.5* 'HalfBath' + 'BsmtFullBath' + 0.5* 'BsmtHalfBath’
- TotalPorchArea' = 'OpenPorchSF' + '3SsnPorch' + 'EnclosedPorch' + 'ScreenPorch’

Missing Values

- PoolQC, Fence, MiscFeature, Alley has more than 80% missing values so we will remove them
- LotFrontage: Filling missing values of street length with street length of neighborhood
- Categorical variables are replace with modal value
- Some categorical variables are filled missing values with ‘None’ because eg Basement is not existent then basement condition value is of no use.

Variable Encoding

- Numerical variables are standardized so that all values are mean centered and have 1 standard deviation
- Categorical variables are dummy encoded/one-hot-encoded
- Concatenated numerical and categorical variables. Now they can be fed to Random Forest Regressor

Feature Selection

- 1. PCA: We tried but did not use PCA for feature selection because PCA provides the principal components which are projection of original variables(in the direction of maximum variance) but we want to use original variables. PCA is typically used for numerical variables only and our data has majority of categorical variables.
- 2. Random Forest Regressor is used for feature importance

Feature Importance: Random Forest

Top 30 features:
Total area of house in square feet variable has highest importance of 0.35 in predicting the house prices.

Next, variable is OverallQual of house having feature importance of 0.32

Next variables shows very low feature importance as compared to above two.

Year Built of House construction, Lot area and GrLivArea