FDS Fall 2017 – Kaggle Project

Kaggle Account: pochiraju

https://www.kaggle.com/pochiraju

Venkata Naga Sai Krishna Abhinay Pochiraju / 1819771
Mani Niharika Racuri / 1819748
Alessandra Griesi / 1578970

1st attempt:

- Linear regression of SalePrice vs GrLivArea

Data Tidying:

- Replacing all the NA values with the mean

Model Selection:

- Linear Model

Kaggle Score: 0.28918

2nd attempt:

- Linear Regression of SalePrice vs OverallQual+YearBuilt+YearRemodAdd+TotalBsmtSF+X1stFlrSF+GrLivArea+FullBath+TotRmsAbvGrd+GarageCars+GarageArea

Data Tidying:

- Replacing all the NA values with the mean

Model Selection:

- Linear Model
- Took all the variables with correlation values > 0.5

Kaggle Score: 0.25407

3rd attempt:

Data Tidying:

- Replacing all the NA values with the mean
- Removed outliers - GrLivArea>4000, LotArea>100000, TotalBsmtSF>6000

Model Selection:

- Linear Model
- Took all the variables with correlation values > 0.5

Kaggle Score: 0.24780

4th attempt:

Data Tidying:

- Combined all the observations (train and test) using rbind()
- Removed outliers - GrLivArea>4000, LotArea>100000, TotalBsmtSF>6000
- Replacing all the NA values (which don't necessarily mean NA, it just means that variable is not present, ex: LotFrontage, No Masonry veener area in the house or Basement:NA means no Basement etc.) with 0
- Removing some errors, like: GarageYrBlt>2018
- Converting all the non-numeric values to numeric
- Replacing all the NA values with the mean
- Splitting the fullDataset into train and test again after data pre-processing

Model Selection:

- Linear Model

Kaggle Score: 0.18538

5th attempt:

Data Tidying:

- Same as above (4th attempt)

Model Selection:

- Linear Model
- Considering only the main variables (Pr values from 0 to 0.001)

Kaggle Score: 0.17573

6th attempt:

Data Tidying:

- Same as above (4th attempt)

Model Selection:

- Random Forest on the same data above – Kaggle Score: 0.14691
- Random Forest for the log of SalePrice and then the prediction is exp(predicted value) – Kaggle Score: 0.14364

Kaggle Score: 0.14364

7th attempt:

Data Tidying:

- Same as above (4th attempt)

Model Selection:

- Xgboost on the same data above

Kaggle Score: 0.13788