MLND_Capstone_Proposal_House_Prices

December 18, 2016

1 Machine Learning Engineer Nanodegree

1.1 Capstone Proposal

1.1.1 House Prices

Hesham Shabana
December 18th, 2016

1.2 Proposal

1.2.1 Domain Background

House prices is a topic that always attract a lot of attention due to the daily deman and the wide range of interseted parties investors, real state agents, taxt estimators, houseowners, and house buyers with this the deman for a reliable model to assist a house price is needed. Traditionaly house price prediction does not take into consideration the wide range of avilable parameters and it also assume indepedndance between these parameters which does not hold in practice.

In Egypt buying a house is ver important decision espicially for young people who are looking to start a family and this decision becomes even harder with thi increase in population, Egypt population estimated to be 93,383,574 with almost 2% growth in the last 4 years, and there is very little research in this area as well as a lack of data. Not to mention that people are following only one rule what is the quare feet price in specific area? Therefore, Using a machine learning algorithm to learn from the past purchasing history will help us to determine and predict the price of the house taking into consideration the attributes that matter the most and this will support any new buyer or a seller to set the right expectation.

This dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

- http://www.doc.ic.ac.uk/~mpd37/theses/2015_beng_aaron-ng.pdf
- https://docs.google.com/viewer?url=patentimages.storage.googleapis.com/pdfs/US6609109.pdf
- https://researcharchive.lincoln.ac.nz/bitstream/handle/10182/5198/House_%20price_%20prediction.pdf
- http://link.springer.com/article/10.1007/s11146-007-9036-8
- https://ww2.amstat.org/publications/jse/v19n3/decock.pdf

1.2.2 Problem Statement

The goal is to analyize historical sales of house prices features and through feature selection, feature engeneering, machine learning find the most relevant set of features that affect the price and which will allow us to perform an accurate prediction for new houses.

To avoid curse of dimentionality given the large number of features included in our database (80 feature) first we need to reduce our vector space by applying hill climbing to conduct feature selection, another proposed technicque to reduce our feature space is to applay feature transformation by leaveraging algorithms like PCA, ICA, Lasso regression. The goal here is to reduce our input space as much as possible to avoid overfitting.

1.2.3 Datasets and Inputs

This dataset describing the sales for houses in Ames, Iowa from 2006 to 2010 which contains 2930 observations and 80 variables, the variable distribution is 23 nominal, 23 ordinal, 14 discrete and 20 continuous.

Next are some statistics regarding the target variable and the input features.

Statistics on the Housing Price variable:

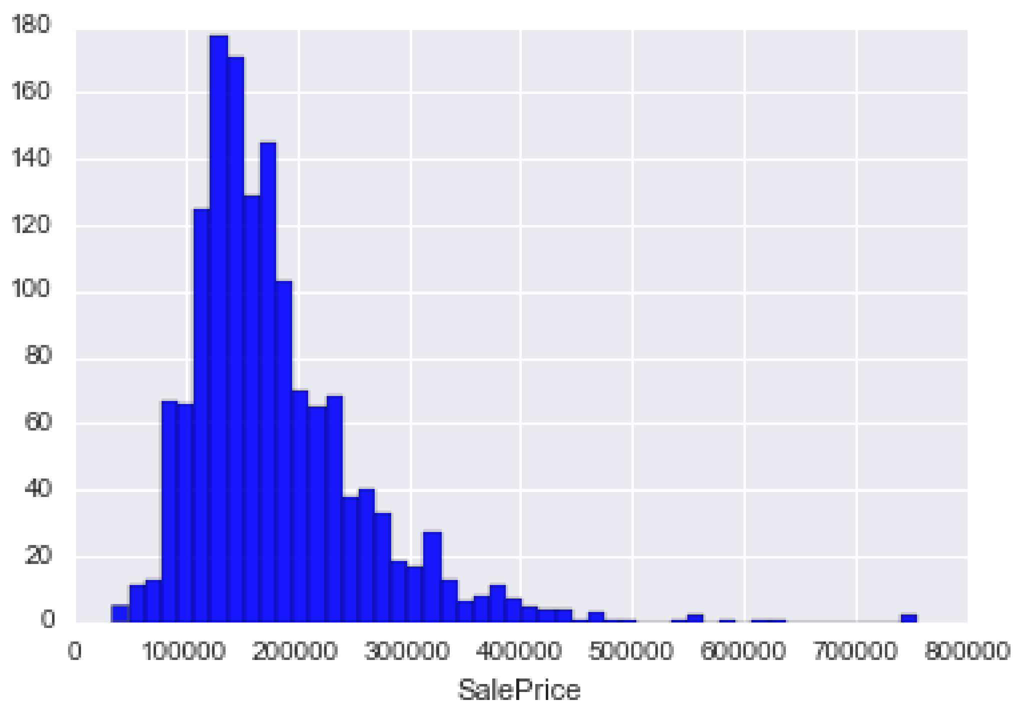
```
In [12]: data['SalePrice'].describe()
```

```
Out[12]: count      1460.000000
         mean     180921.195890
         std       79442.502883
         min       34900.000000
         25%      129975.000000
         50%      163000.000000
         75%      214000.000000
         max      755000.000000
         Name: SalePrice, dtype: float64
```

Housing price distribution

```
In [14]: sns.distplot(data['SalePrice'], kde = False, color = 'b', hist_kws={'alpha': 0.9})
```

```
Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x115b74190>
```

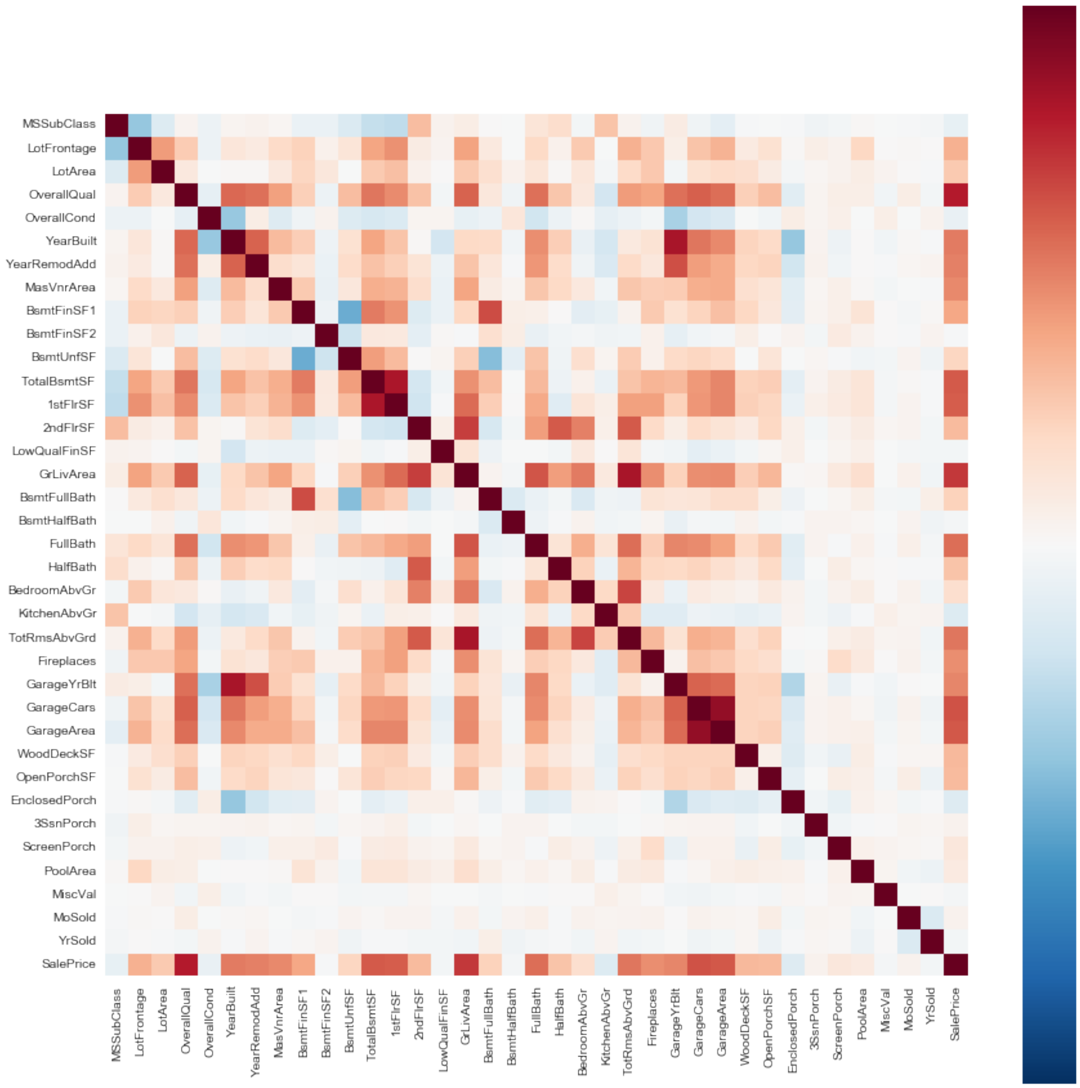


Data features

Numerical features The large number of continuous features (20) and discrete features (14) in this dataset give us various methods to combining and selecting the most relevant features.

```
In [20]: numerical_features_correlation = data.select_dtypes(include = ['float64', 'int64']).iloc[:, 1:
         plt.figure(figsize=(15, 15))
         sns.heatmap(numerical_features_correlation, vmax=1, square=True)
```

```
Out[20]: <matplotlib.axes._subplots.AxesSubplot at 0x11888af90>
```



Categorical features This dataset als has large number of categorical variables (23 nominal, 23 ordinal). They range from 2 to 28 classes with the smallest being STREET (gravel or paved) and the largest being NEIGHBORHOOD (areas within the Ames city limits). In addition, the nominal variables identify various types of dwellings, garages, materials, and environmental conditions while the ordinal variables typically rate various items within the property

```
In [21]: cat_features = data.select_dtypes(include=['object']).columns.values
```

```
In [22]: print cat_features
```

```
['MSZoning' 'Street' 'Alley' 'LotShape' 'LandContour' 'Utilities'
 'LotConfig' 'LandSlope' 'Neighborhood' 'Condition1' 'Condition2'
 'BldgType' 'HouseStyle' 'RoofStyle' 'RoofMatl' 'Exterior1st' 'Exterior2nd'
 'MasVnrType' 'ExterQual' 'ExterCond' 'Foundation' 'BsmtQual' 'BsmtCond'
 'BsmtExposure' 'BsmtFinType1' 'BsmtFinType2' 'Heating' 'HeatingQC'
 'CentralAir' 'Electrical' 'KitchenQual' 'Functional' 'FireplaceQu'
 'GarageType' 'GarageFinish' 'GarageQual' 'GarageCond' 'PavedDrive'
 'PoolQC' 'Fence' 'MiscFeature' 'SaleType' 'SaleCondition']
```

1.2.4 Solution Statement

A solution for this problem is to take the input features and make them suitable to be used by a machine learning algorithms, in addition, to do feature selection and engineering to enhance the performance of the model, at the end we shoul have a list of the most relevant features and an accurate model that could be use to predict new house prices.

1.2.5 Benchmark Model

As a benchmark the variation in residential sales price can be explained by simply taking into consideration the neighborhood and total square footage of the house. On the other extreme, constructe a model with 36 features explains 92% of the variation in sales.

1.2.6 Evaluation Metrics

To assesst the model performance the below matrics will be used: * Confusion matrix (presision, recal and F1) * R^2 score: to evaluate the regression model * Bias: positive values indicate the model tends to overestimate price (on average) while negative values indicate the model tends to underestimate price. * Maximum Deviation: identifies the worst prediction made in the validation data set. * Mean Absolute Deviation: the average error.

1.2.7 Project Design

To solve this problem the following steps shall be executed:

- Process the data
- Deal with outliers and null data
- Transform categorical features
- Standardize numerical features
- Decide weather to split the data to training and testing sets or to user cross-validation depending of the size of the data
- Perform Feature selection by using hill climbing technique
- Feature transformation by using PCA algorithm, Lasso regression
- Feature engineering
- Transform the testing data feature space
- Train our model using the below algorithms:
- Logistic regression
- Boosting
- random forest
- Evaluate the model performance
- Applay our model to the testing data