Machine Learning Engineer Nanodegree

Capstone Proposal

Santosh Narayan
February 13th, 2018

Proposal

Domain Background

This project is aimed at predicting house prices. The Ames Housing dataset was compiled by Dean De Cock.

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

The Ames Housing data set has about 80 variables that directly relate to property sales. This dataset focuses on both quantitative variables and qualitative variables of the project. The variables are the perfect representation of the questions any typical buyer would ask before buying a house/property. Example include: When was the house built? Does it have a basement? How many bathrooms does the house have? The dataset has information that can answer a typical buyers question. The datset represents the sale of individual residential property in Ames, Iowa from 2006 to 2010. 10. The data set contains 2930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing home values.

Property/Real Estate is a very widely used measure of a person's wealth. A home is the probably the biggest investment anyone makes. It is a field with relatively low barriers to entry, anyone can participate at a level of the purchasing power they have. Flipping houses has become an occupation for quite a few. By knowing the true market value, one can help arbitrage and earn significant profits.

Hence, I have selected the Ames, Iowa dataset of housing sales prices to analyze and develop a model to predict housing prices. The project is taken from Kaggle competition. In the future, using this model as a base, I would like to be able to predict at my local area.

I would also like to cite the following research paper: https://smartech.gatech.edu/bitstream/handle/.../Corsini_Kenneth_R_200912_mast.pdf which discusses the housing prices in relation to other external factors such as employment rate, interest rates, New construction, Consumer Price Index etc.

Problem Statement

https://en.wikipedia.org/wiki/Kaggle

I plan to use regression techniques such as random forest and gradient boosting.

Inputs are the 79 variables and out put is to predict the price of the house.

The opportunity involves in using the data set provided to predict the housing price. The goal is to predict the sale price variable of each house. There are so many different variables in the dataset that come into play when predicting a house price. Every solution and analysis can be unique depending upon the user. I look forward to using my skills and knowledge in this domain to come up with a prediction.

https://ww2.amstat.org/publications/jse/v19n3/decock.pdf

Datasets and Inputs

The dataset is provided on Kaggle competition website. They are available to download by everyone.

https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

```
Data set contains information from the Ames Assessor's Office used in
computing assessed values for individual residential properties sold in Ames,
IA from 2006 to 2010
```

The data set given contains 1460 observation in the train file and 1459 on the test file and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing home values. The 14 discrete variables typically quantify the number of items occurring within the house. Most are specifically focused on the number of kitchens, bedrooms, and bathrooms (full and half) located in the basement and above grade (ground) living areas of the home. Additionally, the garage capacity and construction/remodeling dates are also recorded. There are a large number of categorical variables (23 nominal, 23 ordinal) associated with this data set. 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). The nominal variables typically identify various types of dwellings, garages,

materials, and environmental conditions while the ordinal variables typically rate various items within the property.

data_description.txt provides the detailed description of the data.

Here's a brief version of what you'll find in the data description file. The below variables are the list provided on Kaggle by the data provider.

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition

- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

Solution Statement

I plan to use the various regression techniques such as XGBoost regressor, Random forest regressor, SVM regressor etc. to arrive at the solution. Kaggle has provided the ‘train’ and ‘test’ datasets. I plan to train the algorithm first and then be able to verify the effectiveness of the algorithm using the test dataset.

The final solution algorithm is such that, given the values for all the features, the algorithm will predict the price of the house.

Benchmark Model

I plan to use to random forest as the benchmark model.

Evaluation Metrics

The Evaluation metric provided by Kaggle for this problem is as below

https://www.kaggle.com/c/house-prices-advanced-regression-techniques#evaluation

The submissions will be evaluated using Root Mean Squared Error(RMSE) between the predicted value and the observed value and the logarithm of the observed sales price.

Further, as this is a Kaggle competition project, I would be using the leaderboard score as my evaluation. Prediction results will be evaluated by comparing the predicted values and the actual values provided by the assessor's office.

Project Design

The steps I plan to take in the project design are below. These steps are very similar to what I have been exposed as part of my assignments in Udacity.

Data Analysis

As part of this process I will first seek to understand the variables used in the data and what they mean. I will also do some research outside of the project to ensure that my understanding of the variable is exactly what the variables imply in the value provides. This data analysis will also provide me insights into the data pre-processing that needs to occur.

Data Visualization

This is a very important part of the project where I will be visualizing the training data's target variable. I also plan on understand visually the relationships between other variables in the project to see how they relate to each other. This will help me if I must introduce any feature engineering in the process of solving this problem

Data Preprocessing

Here I will use the insights gained from data analysis and visualization to clean and re-structure the data. I will use the strategies that have been taught in the past to handle missing data. Outliers in the

data can be seen from data visualization. They can be handled several ways, some of the methods are by removing them or replacing them with mean values etc. This step will also take care of the categorical variables in the data and encoding them

Data Split and Modeling

Although Kaggle has provided a test set, I would like to gauge the accuracy on my own test set (by splitting the data into Training/Test (75/25 split)). I plan on doing a benchmark as mentioned in the bench mark section to get the benchmark accuracy. After data pre-processing and feature engineering I will train the data on the model and get a second accuracy score. I am confident that this accuracy score of the tuned data sets will be higher than the benchmark score.