Machine Learning Engineer Nanodegree

Capstone Proposal

Huang Cheng, Lin

August 22nd, 2018

Proposal

Domain Background

A house value is simply more than location and square footage. Like the features that make up a person, an educated party would want to know all aspects that give a house its value.

For my personal motivation, when I started experimenting with machine learning, I wanted to come up with an application that would solve a real-world problem but would not be too complicated to implement. I also wanted to practice working with regression algorithms. So I started looking for a problem worth solving. Here's what I came up with.

Problem Statement

The goal of the project is applying basic machine learning concepts on data collected for housing prices to predict the selling price of a new home. Since this is a typical regression problem, we will use a couple of regression models to predict housing prices and find the best one.

Datasets and Inputs

The dataset is obtained from Kaggle. The dataset has 1461 instances and 81 attributes including target variable. The 81 attributes are described as follows:

- 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

Since the target variable ’SalePrice’ in this project is continuous and it will be estimated from the other variables, we will use regression analysis for this problem. There are a few regression models available in scikit-learn like Decision Tree Regression, Logistic Regression, Stochastic Gradient Descent Regression, etc. We will use a couple of regression models to predict the house prices and find out the best model.

Benchmark Model

Linear regression comes with a set of implicit assumptions and is not the best model for every situation. In this project, we will use linear regression as our benchmark model.

Evaluation Metrics

In this project, we will use the coefficient of determination, R^2, to quantify the model's performance. The coefficient of determination for a model is a useful statistic in regression analysis, as it often describes how "good" that model is at making predictions. The formula of R^2 is given by

$$R^2 = 1 - \frac{SSE}{SST}$$

Where SSE is the sum of squared errors of our regression model

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

And SST is the sum of squared errors of our baseline model.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

The values for R^2 range from 0 to 1, which captures the percentage of squared correlation between the predicted and actual values of the target variable. A model with an R^2 of 0 is no better than a model that always predicts the mean of the target variable, whereas a model with an R^2 of 1 perfectly predicts the target variable. Any value between 0 and 1 indicates what percentage of the target variable, using this model, can be explained by the features.

Project Design

First, In order to understand the data, we will make a cursory investigation about the dataset and visualize our data to find the degree of correlations between predictors and target variable.

Before we use the data to develop a model, we need to make sure that it is in a useful scale, format and even that meaningful features are included. This step is

typically known as preprocessing. In this step, we will be handling the missing data, outlier and using methods like PCA to reduce dimension.

When the data is considered good enough to be used, we will choose a couple of supervised learning models that are currently available in scikit-learn and train them with the dataset.

To properly evaluate the performance of each model that has been chosen, we will create a training and predicting pipeline that allows we to quickly and effectively train models using various sizes of training data and perform predictions on the testing data.

Final, based on the evaluation, we will choose the best model and fine tune it using grid search method.