Machine Learning Engineer Nanodegree

Capstone Proposal

Xing Lu
November 2nd, 2018

Proposal

Domain Background

Housing market is very competitive in a growing economy background. Being able to estimate selling price of a home is very useful for both customers and listing parties. It is interesting to know what influences price negotiations. Supervised learning is one of the most promising field in machine learning that is currently used in real world, and it will be used in this study to predict housing prices.

Problem Statement

Given a large residential home dataset for Ames, Iowa with 79 explanatory variables (continuous or discrete), the final price of each home will be predicted using supervised learning models and deep learning model.

Datasets and Inputs

Datasets for the problem can be retrieved from Kaggle[1] through https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data.

The datasets contains two .csv files, train.csv and test.csv with information of homes to make a prediction.

Id: The id of a home.

79 variables from col. 2 to col. 80: 79 different features of a home.

SalePrice: The target variable. Not presented in test.csv.

The size of train.csv is 1460 x 81 and the size of test.csv is 1459 x 80. There are 1460 homes in train.csv and 1459 homes in test.csv.

Solution Statement

We want to understand the relationship between the 79 features and the housing price. Skewed continuous features will be transformed. Numerical features will be normalized to ensure each feature is treated equally. Non-numeric features will be one-hot encoded. After preprocessing of the data we will build models to make prediction. We will try logistic regression, support vector machine and Adaboost. Grid Search will be used to tune parameters. We will also build a deep learning model in Keras to make prediction. The exact neural network will be decided while building and testing the model.

Benchmark Model

A benchmark model for this project would be the best Kaggle score since this is a Kaggle competition. The current best submission has a 0.06628 root mean squared logarithmic error (the lower the better). The best Kaggle score is calculated based on 50% of the test model, and we don't know which 50% is used. Thus, we cannot compare between our model and the benchmark model before submission (limit to 5 per day). Because there is no sale price provided for the test set, a part of the training set (last 260 entries) will be used as the test set to evaluate and compare the performance between the different models I build for this proposal using the evaluation metric described below. The results from the model with the best performance will be submitted to Kaggle for official score.

Evaluation Metrics

As per Kaggle, models will be evaluated on Root Mean Squared Error[2] (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. Taking logs means that errors in predicting expensive houses and

cheap house will affect the result equally. Thus, RMSE will be the evaluation metric for this proposal.

Project Design

Preprocessing data: Transform skewed continuous feature. Normalize numerical features. One-hot encode non-numeric features.

Shuffle and split data: Create cross validation set from the training data. Shufflesplit data in train.csv.

Model application: Build supervised models (logistic regression, support vector machine, Adaboost). Tune parameters using grid search. Build deep learning model in Keras. Fit, train, and evaluate models. Choose results from the best model with the lowest RMSE to submit.

References

[1] Kaggle: House Prices: Advanced Regression Techniques. https://www.kaggle.com/c/house-prices-advanced-regression-techniques#evaluation

[2] Root-mean-square deviation Wikipedia. https://en.wikipedia.org/wiki/Root-mean-square_deviation