Machine Learning Engineer Nanodegree

Capstone Proposal

Lucas Rafagnin

December 13st, 2019

Domain Background

The project is a Kaggle competition[1] called House Prices: Advanced Regression Techniques. This dataset provide a great opportunity to practice some machine learning alghoritms and data science methods. The problem to be solved in this project is a very important theme in real estate world, I say this because I work in a company[2] that has web platforms that manage real estate in Brazil and believe that solutions like this would help the business.

Problem Statement

In this project I will predict the house price based on some explanatory variables describing aspects of residential homes in Ames, Iowa.

Datasets and Inputs

The data[3] is provided by the Kaggle competition and has 79 explanatory

variables with information about the residential homes in Ames, Iowa. The dataset has 2 parts:

- Training data included 1460 houses accompanied by 79 attributes and the sales price for each house.
- Testing data included 1459 houses with the same 79 attributes, but sales price was not included.
  Of the 79 variables provided, 51 were categorical and 28 were continuous.

Solution Statement

Firstly I will examine the features and their correlations, then process the data, filling the missing and nullable values, select features and apply mathematical functions or regression methods to trains the model. After that, with the model in hands the final performance can be evaluated by kaggle public leaderboard.

Benchmark Model

For the benchmark model a comparison with the leaderboard will be used. In addition I will also use this submission (which got a score of 0.11420) as a benchmark of my model.

Evaluation Metrics

It is project goal is to predict the sales price for each house, and the metric is evaluated on Root-Mean-Squared-Error (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 houses will affect the result equally).

Project Design

Language: Python3
Some libraries: Numpy, Pandas, Matplot, Seaborn and Sklearn

Initially understand the dataset better, exploring and discovering each of its attributes, after that I will follow the workflow:

1. Explore data:

Exploring the dataset possibilities and analyzing the correlations between its features. So here are some steps:

- Basic stats via boxplot (min, max, std, median…)
- Frequency via histogram
- Data types (number, text, object, etc.)
- Correlation analysis

2. Clean data:

Data need to be cleaned and transformed before trying different ML algorithms. Therefore some methods are possible:

- Handle missing data;

- Remove skewness;
- Outlier analysis;

3. Prepare Data:

Create the traning and test sets using proper sampling methods, e.g., random vs. stratified.

4. Feature treatments:

With 79 features the steps like feature engineering and feature transformation are important:

- Separate the predictors and labels
- Convert text attributes into numerical attributes
- Apply feature engineering to reduce the number of features
- Feature scalling (normalization or standardization)

5. Model selection:

The objective here is to analyze and compare the multiple regression models, some models are important to that goal:

- LinearRegression
- GradientBoostingRegressor
- XGBoost
- LightGBM

Lastly I will apply the final model to the data and submitting the results

to Kaggle for check the competition score.

*Remembering that this step by step is a draft and can undergo changes as needed during development

Reference

[1] kaggle.com. House prices: Advanced regression techniques. **https://www.kaggle.com/c/house-prices-advanced-regression-techniques**.

[2] grupozap.com. Grupo ZAP. **https://www.grupozap.com**.

[3] kaggle.com. House prices: Advanced regression techniques. **https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data**.