Machine Learning Engineer Nanodegree

Capstone Proposal

Aditya Rao

26-03-2018

Proposal

In this project I am trying to solve the Kaggle challenge of "Zillow Prize: Zillow's Home Value Prediction (Zestimate)". The goal of this project is to predict the price for three counties (Los Angeles, Orange and Ventura, California). The training data comprises over 58 features for each house.

Domain Background

Zillow, is an online real estate database company Zillow has data on 110 million homes across the United States, not just those homes currently for sale. A home is often the largest and most expensive purchase a person makes in his or her lifetime. Ensuring homeowners have a trusted way to monitor this asset is incredibly important. The Zestimate was created to give consumers as much information as possible about homes and the housing market, marking the first-time consumers had access to this type of home value information at no cost.

Problem Statement

The goal of this project is to predict the price of new properties that are going to be sold in Los Angeles, Orange and Ventura, California. This is a regression problem where the price of the property is evaluated based on its features. In this Kaggle project we are given 6 Million Records and Zillow Prize, a competition with a one-million-dollar grand prize, is challenging the data science community to help push the accuracy of the Zestimate even further. Winning algorithms stand to impact the home values of 110M homes across the U.S.

Datasets and Inputs

This project will use the dataset provided by Zillow on their Kaggle competition page. The dataset consists of both 2016 and 2017 CSV files of properties with 58 features and around 6 Million records. Some of features include House Area, Number of Bedrooms and Bathrooms, Year of built, Pool size, Yard size, etc.

The dataset will be used to train an algorithm to estimate the price of the house, after which the prediction will be compared Kaggle leader board.

Solution Statement

I plan to use multiple Supervised algorithms from the SkLearn Library which will help me in providing better insights on the data along the way, but I intend to use a Neural Network, as there are many feature labels in the dataset. The Deep Learning model will be built with TensorFlow and Kears.

Benchmark Model

A major benchmark model will be one of the submission on the Kaggle leader board , Here the model used is a simple Linear Regression which got a score of "(LB ~ 0.0666)"

Evaluation Metrics

Submissions are evaluated on Mean Absolute Error between the predicted log error and the actual log error. The log error is defined as

$$Logerror = \log(Zestimate) - \log(SalePrice)$$

and it is recorded in the transactions training data. If a transaction didn't happen for a property during that period of time, that row is ignored and not counted in the calculation of MAE.

Project Design

Programming Language: Python 3

Libraries: Numpy, Pandas, Matplotlib, Seaborn, TensorFlow, Keras.

Workflow:

- Provide basic analysis on the data with visualization.
- Cleaning of the data, Handling categorical Attributes, Feature Scaling and Splitting data from Training and Testing.
- Providing a baseline model
- Applying various models make a better understanding
- Fine tuning the model with Grid Search/Ensemble Methods/Tuning Hyperparameters
- Applying the final Model on to the data and submitting the results to Kaggle for a more refined score.