Machine Learning Engineer Nanodegree

Capstone Proposal

Swetha Ganapathi Raman

Mar 8th 2017

The capstone project I am choosing for MLND is a current challenge in Kaggle - Two Sigma Connect : Rental Listing Inquiries
https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries

Domain Background

For this project, I am interested in predicting the popularity of a rental listing.

Real state companies like Redfin have become very popular in the recent years, mainly due to using data science/Machine learning/predictive analytics techniques to accurately estimate the price of a home.

It would be very interesting to use data science/ ML techniques for the rental market as well. Renthop is a rental listing company that does this to rank rental listings by quality. This makes it easier for people looking to rent homes, to find a high quality rental listing quickly instead of having to sift through all the listings, especially in hot rental markets like San francisco bay area, NYC etc.

Citations:

Housing Value Forecasting Based on Machine Learning Methods

Jingyi Mu,1 Fang Wu,2 and Aihua Zhang3

https://www.hindawi.com/journals/aaa/2014/648047/

Problem Statement

This problem features rental listing data from Renthop. The objective is to predict the popularity of a new listing (high interest, medium interest or low interest) based on the listing content like text description, number of bedrooms, price and other features. The dataset features numerical, categorical, text and image data as input fields. The target variable, interest_level is defined by the number of inquiries a listing has in the duration that the listing was live on the site.

As stated by Renthop, “This prediction will help Renthop handle fraud control, identify potential listing quality issues and allow owners and agents to better understand renter’s needs and preferences.”

I intend to feed these inputs to a couple of Supervised classifier algorithms to see which produces the most accurate output class (High, medium, low), measured by multi class log loss.

Dataset and inputs

Data source for the problem is provided by Renthop.
https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries/data

The dataset has the following files:
Train .json - the training set
Test.json - the test set
Images_sample.zip - listing images organized by listing_id (a sample of 100 listings)

Solution statement

A solution to this problem, would be to use a set of the input features, run them through a couple of supervised classifier models, to produce the predicted output classes. The model which has a lower multi class log loss would be a good solution.

Benchmark model

For the benchmark model, I would use a model that always predicted that the listing has a ‘low’ interest, i.e. the output class is ‘low’

Evaluation metrics

The model will be evaluated using multi class logarithmic loss.

$$Log\ loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} Yij\ (Pij)$$

Where N = number of listings in the test set
M = number of class labels (3 classes)
Log is the natural logarithm
Yij =1 if observation i belongs to class j and 0 otherwise
Pij = Predicted probability that observation i belongs to class j

Multi class log loss defined by Kaggle
https://www.kaggle.com/wiki/MultiClassLogLoss

Project design

My project design would be as follows:

1. Data exploration
    a. Preliminary investigation of data
2. Preprocessing
    a. Look for any invalid or missing entries or outliers
    b. Transform any non numeric features to numeric as needed
    c. Create any new features
    d. Split data into training and test set
3. Modeling
    a. Generate 3 models for comparison - One benchmark model, two other supervised classifiers
    b. Compare performance of all the 3 models
    c. Choose the best model and optimize its parameters using gridsearch
    d. Report optimized model's performance compared with the benchmark