Machine Learning Engineer Nanodegree

Capstone Proposal

Leon Paul
November 10th, 2017

Domain Background

Regression analysis has been in wide use since the start of the 19th century beginning with the application of the Ordinary Least Squares method, by Legendre and Gauss, to determine the orbits of comets around of Sun based upon astronomical observations [1]. Regression methods have seen widespread usage within two key areas, namely **Forecasting** and **Optimization,** with a view to extract valuable insights and drive business decisions.

Forecasting the sale price of a house given a variety of information pertaining to it and it's neighbourhood is a prime importance in the present, with companies like Zillow and ZipRealty using empirical data to forecast estimates of property prices for their customers in real time. The ability to create models which can make these predictions with as high an accuracy as possible is what gives these organizations a competitive edge. Some previous applications of regression applications in predicting/modelling house prices have been seen in "Modelling Home Prices Using Realtor Data, 2008, Iain Pardoe" [5] and "House Price Prediction: Hedonic Price Model vs. Artificial Neural Network, 2004, Visit Limsombunchai". [6] I plan to use these academic papers as guides for my capstone.

I have an interest in forecasting trends or predictions within the real estate, financial and retail sectors and as a result I decided to work within this area for my Capstone.

Problem Statement

For my Capstone, I will be working with the Ames Housing dataset [2]. This dataset consists of 79 explanatory features describing a variety of informative details about residential houses sold in the city of Ames, Iowa between 2006 and 2010. The datasets have been made available through a Kaggle competition "**House Prices: Advanced Regression Techniques**" [3].

The problem requires us to predict the Sale Price of a house given the set of 79 explanatory features. The solution submitted to Kaggle will be evaluated by them using the Root Mean Square Error between the logarithm of the predicted value and the logarithm of the observed sales price. Taking the logarithm will normalize the errors ensuring that the prediction errors for expensive and cheap houses will be weigh equally on the model's evaluation.

$$RMSE = \sqrt{E((\log(\hat{\theta}) - \log(\theta))^2)}$$

$\theta := Actual\ Sale\ Price$

$\hat{\theta} := Predicted\ Sale\ Price$

Datasets and Inputs

The dataset in question is available freely from a number of sources. I was able to download the training and testing datasets from the Kaggle Competition webpage [3]. The dataset consists of 79 input features which are a combination of categorical and numeric data types. The response variable is a continuous value that represents the Sale Price of the houses. The datasets have the following characteristics:

TRAINING DATASET:

- **1460 rows**
- **79 input features**
- **1 response feature ~ Sale Price**
- **43 Categorical features**
- **36 Numerical features**

TESTING DATASET:

- **1459 rows**
- **79 input features**
- **43 Categorical features**
- **36 Numerical features**

The response variable will be modelled using the given 79 input features. The key area of focus will be on Feature engineering and regression models that minimize the RMSE score.

Solution Statement

The essential approach would be to perform Exploratory Data Analysis in order to determine the distribution of the dataset's numerical features, determine the number features having missing values as well as the percentage of missing values. Following this, the next step would be to use feature engineering to transform, create or drop features so that any feature that does not contribute significantly to a model's information gain would not be used and help reduce the dataset's dimensionality. Lastly, a regression model would be run using the response variable as a target label and a combination of the pre-processed predictors to train the model. Some of the models I aim to try are Multivariate Linear Regression, Neural Nets and Decision Trees. Cross-Validation and ensemble methods like Random Forests or Gradient Boosting would be used to improve the performance and reduce overfitting.

Benchmark Model

The benchmark model would serve as an initial result that every subsequent model should aim to beat. It will also provide a relative measure of how my prediction has improved. As an initial benchmark, I will select a set of features that according to my knowledge would be influential in deciding in the Sale Price of a house, for e.g. square foot area, condition of house, presence of a garage/driveway, number of bathrooms etc. These predictors would be fed into a linear regression model which would serve as the Benchmark for any future models.

In addition, since the dataset is in use in a Kaggle competition, I also plan to use the results from other teams to benchmark my model's relative performance.

Evaluation Metrics

The Evaluation metric used for this problem would be the logarithmic Root Mean Square Error (RMSE). RMSE is defined as "*measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed*" [4]. It is represented as follows:

$$RMSE = \sqrt{E((\hat{\theta} - \theta)^2)}$$

$\theta := Actual\ Feature\ Value$

$\hat{\theta} := Predicted\ Feature\ Value$

The logarithmic RMSE is calculated simply by using the logarithms of the predicted and acual values, as shown below:

$$RMSE = \sqrt{E((\log(\hat{\theta}) - \log(\theta))^2)}$$

This serves to normalize the results ensures that errors from expensive or cheap houses are given equal weightage while improving the model performance. Since the Sale price values are huge; in the region of $100,000 to $1,000,000; the errors between actual and predicted values could be very large and taking the logarithm of these values ensures that such large errors are not penalized.

Project Design

I plan to begin by performing a detailed Exploratory Data Analysis (EDA) on the dataset to understand how the various features are distributed and determine the percentage of missing features.

This EDA would be followed up by feature engineering, imputation of missing values and feature transformations. The categorical variables would have to be encoded into numeric variables in order to fit them into models like Linear Regressions.

The third step would involve setting up the benchmark model viz. Linear Regression using a subset of features selected using our intuition and domain knowledge. The process has been discussed in detail earlier.

The next step would be to determine an optimum algorithm for the final model and pre-process the datasets for that model. My choice would be with ensemble methods like Gradient Boosted Models or Random Forests. I would use a mix of ensemble learning and k-Fold Cross Validation to tune the model and optimize it for the pre-processed data.

With the model ready, it's parameters tuned and the data pre-processed I will then train my final model on the training data and generate predictions using the testing data set.

The final step would be to calculate the Logarithmic RMSE score for the model in order to determine how much better it is over the benchmark model.

References:

[1] https://en.wikipedia.org/wiki/Regression_analysis
[2] https://www.openintro.org/stat/data/ames.php
[3] https://www.kaggle.com/c/house-prices-advanced-regression-techniques#description
[4] https://en.wikipedia.org/wiki/Root-mean-square_deviation

[5] https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=8&cad=rja&uact=8&ved=0ahUKEwjx5o2d2LTXAhVWwWMKHWwWDeAQFghbMAc&url=http%3A%2F%2Fww2.amstat.org%2Fpublications%2Fjse%2Fv16n2%2Fdatasets.pardoe.pdf&usg=AOvVaw2On9ZvayT59Lr5EjoDn1Qi
[6] https://ageconsearch.umn.edu/bitstream/.../2004-9-house%20price%20prediction.pdf