Machine Learning Engineer Nanodegree

Capstone Proposal: Sberbank Russian Housing Market

Pramod Rao

20-July-2017

1. Domain Background

A lot of people working in data-related fields heard of Kaggle: a platform for running data science competitions. Currently Kaggle features more than 600K data scientists and dozens of well-known companies. A company defines its problem, quality metrics, publishes a dataset which may help solve it — and participants show off their creativity in solving company's problem.

Housing costs demand a significant investment from both consumers and developers. And when it comes to planning a budget—whether personal or corporate—the last thing anyone needs is uncertainty about one of their biggets expenses.

This project derived from Kaggle competition is brought by Sberbank, Russia's oldest and largest bank, helps their customers by making predictions about realty prices so renters, developers, and lenders are more confident when they sign a lease or purchase a building.

2. Problem Statement

This project is based on the Kaggle competition[1]: "**Sberbank Russian Housing Market".** The aims to provide a solution by developing models using machine learning with deep learning to predict the reality prices. The project is largely rely large spectrum of features from the data set to predict the reality price.

3. Datasets and Inputs

The project will use the data provided by the Sberbank on the Kaggle competition forum. The dataset is classified as following:

- train.csv, test.csv: information about individual transactions. The rows are indexed by the "id" field, which refers to individual transactions (particular properties might appear more than once, in separate transactions). These files also include supplementary information about the local area of each property.
- macro.csv: data on Russia's macroeconomy and financial sector (could be joined to the train and test sets on the "timestamp" column)
- sample_submission.csv: an example submission file in the correct format
- data_dictionary.txt: explanations of the fields available in the other data files

The dataset can be downloaded from: https://www.kaggle.com/c/sberbank-russian-housing-market/data by agreeing to the rules of the competition.

---

[1] https://www.kaggle.com/c/sberbank-russian-housing-market

4. Solution Statement

In order to predict the sale price of the real estate property, I will use Ensemble methods such as Random Forest, XGBoost along with Deep Learning. The dataset comprising nearly 300 features would be categorized into different classes and feature engineered accordingly. The primary goal is estimate the price with minimum Root Mean Square Log Error (RMSLE).

5. Benchmark Model

As the project is derived from Kaggle, the Leaderboard models could serve as the benchmark model. The solution from the project could be submitted and compared with the best Leaderboard scores in order to evaluate.

6. Evaluation Metrics

Kaggle platform requires a company that runs the competition, define a transparent clear metric that participants compete on. Sberbank competition features Root Mean Square Log Error as a Metric. RMSLE is measure the ratio between the actual and the predicted value and is defined as:

$$log(pi+1)-log(ai+1) \text{ that is equivalent to } log((pi+1)/(ai+1))$$

where p is the predicted value and a is the actual value.

RMSLE gives more weightage to percentage difference than difference between the numbers. This is metric is fixed and cannot be altered from the competition perspective. Nevertheless, I feel the metric chosen is apt for the problem since RMSLE doesn't penalize large differences when dealing with huge numbers. In this project, the reality prices are relatively huge numbers and more than the difference in cost of actual and predicted sale price value, the percentage difference between the two could be a better scale to evaluate.

7. Project Design

7.1. Programming Language and Libraries

- Python 3
- Scikit-learn: machine learning library based on python
- Kearas: a high-level neural networks API, written in Python and capable of running on top of either TensorFlow or Theano.
- Tensorflow: Open source software for deep learning

7.2. Project Flow:

- Data exploration: Visualize and inspect dataset. Inspection includes check for outliers, skewness and correlation.
- Feature Engineering: Discard the outliers, transform the data to remove skewness and remove variables causing multicollinearity.
- Model training and tuning: Use ensemble methods and neural networks in order to design a model for prediction. As there are wide categories of features, develop multiple models using algorithms such as MLP, XGBoost etc., for different types and finally combine them.

- Results validation: we combine predictions of different models using a linear regression, observe the results and prove their statistical significance.