A Neural Network Approach to Home Price Predictions
Hongtao Sun and Ji Hoon Andy Kim
{s3sunht, jkim4223}@stanford.edu

Motivation
Predicting real estate markets is key to understanding the global economy. There is extensive literature on building models to better predict home prices such as employing PCA, LASSO/Ridge Regression, Random Forests or XGBoost. However, we propose using a feed-forward neural network approach to the home price prediction model and compare its performance against the baseline models.

Dataset
We have two datasets HOUSE and MACRO from Kaggle competition: Sberbank Russian Housing Market

HOUSE contains 30470 records of house information.
• Target variable: price_doc.
• Building features (11 features)
• Neighborhood features (279 features)
• Noisy and sparse with missing values in 51 features have missing values.

MACRO contains macroeconomic indicators from 2010 to 2016 with 2484 data points and 100 features.

To evaluate our models, we use the Root Mean Squared Logarithmic Error (RMSLE) metric

\[ \text{RMSLE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \log(y^{(i)} - \hat{y}^{(i)})^2} \]

The best performing models have the lowest RMSLE values

Models
Baseline Models
• SLR for variable selection
• PCA
• LASSO/Ridge Regression
• Optimal lambda empirically chosen
• Regression Forests
• XGBoost Gradient Boosting
• Parameters also empirically chosen through greedy minimization of RMSLE

Feed-Forward Neural Network
• 3-layer Neural Network with 2 hidden layers
• Using all features including the macroeconomic features
• 25 and 15 neurons respectively
• 40 Epochs

Results
We take the dataset and split it 70-30. We use 70% for training the predictive models and 30% for cross validation. This leaves us with 21, 329 training examples.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>0.5196</td>
</tr>
<tr>
<td>Ridge</td>
<td>0.5068</td>
</tr>
<tr>
<td>LASSO</td>
<td>0.5031</td>
</tr>
<tr>
<td>Regression Forest</td>
<td>0.4607</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.4301</td>
</tr>
<tr>
<td>Neural Network</td>
<td>0.2544</td>
</tr>
</tbody>
</table>

Table 1: Home price prediction models and the respective RMSLE metrics

Evaluating the benchmark models against the hold-out cross validation set, we achieve the best RMSLE scores with the XGBoost gradient boosting algorithm.

However, with the feed-forward neural network model we specified, we are able to achieve vastly better result than all the benchmark models. The confirms with our expectation that feed-forward neural network models are viable in solving tasks like predicting housing prices.

Further Work
• Perform a more statistically sound method of imputing data.
• Expand computational power to include more computing cores to perform neural network training en masse.
• Run the feed-forward neural network over more training data and across more epochs.
• Generate an ensemble between neural network models and gradient boosting models.

Data Manipulation
As inputs to the model, one-hot encoding is used to incorporate the categorical variables to feed into our models.

References