Airbnb Price Estimation

Liubov Nikolenko liubov@stanford.edu
Hoormazd Rezaei hoormazd@stanford.edu
Pouya Rezazadeh pouyar@stanford.edu

Stanford University

CS 229: Machine Learning
Autumn 2018

Introduction

- **Why?** Evaluating the price of a listing on Airbnb is challenging for the owner as well as the customers who have minimal knowledge of an optimal price. This project aims to provide a price prediction model to help solve this challenge.
- **What?** Several models have been studied ranging from linear regression to tree-based models and neural nets. To boost the models' performance several feature selection schemas have been explored.
- **Results:** Using feature selection, hyperparameter tuning, and a variety of models, the R^2 of the prediction was improved from a negative value to 69% for the best model.

Dataset

- **Dataset:** Public Airbnb dataset for New York City[1]
- **Data Description:** ~50,000 examples (listings) with 96 features from owner information to property characteristics such as number of rooms and bathrooms as well as geographic coordinates of the listings
- **Labels:** Price of the listing (ground truth)—also included in the data (figure 1 shows the geographic spread of the labelled data points)

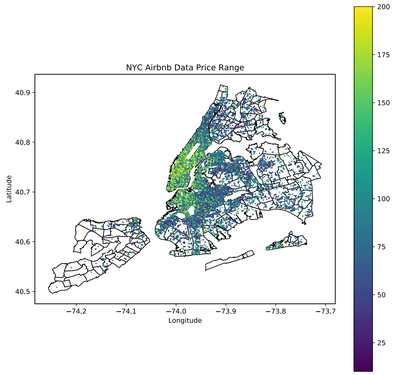

Fig 1: Geographic spread of price labels (with filtered outliers)

Feature Description

- The raw dataset included 96 features including categorical and ones with too many missing entries
- The incomplete features were removed and the categorical ones were transformed into one-hot vectors
- Raw text of the listings reviews was analyzed using TextBlob[4] module
- Data preprocessing resulted in 764 features which were trimmed down using the following feature analysis methods:
  - Manual feature selection
  - P-value feature importance analysis
  - Lasso cross-validation feature importance analysis
- The resulting R^2 values for the reduced feature sets are included in figure 2

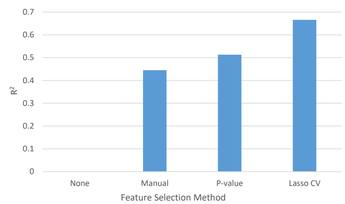

Fig 2: Performance improvements in the model due to different feature selection schemas

Models

- **Ridge Regression** with objective function $J(\theta) = ||y - X\theta||^2 + \gamma||\theta||^2$
- Classifying the data points using **k-means clustering** into one of the groups ($c_j^{(i)} = \arg\min_j ||x^{(i)} - \mu_j||_2$, where j is the index of the group) and using **Ridge Regression** trained on that specific group
- **Support Vector Regression**[3] with **RBF kernel** $K(x,z) = \exp\left(-\frac{||x-z||^2}{2\sigma^2}\right)$

$$\min_{\omega,b,\xi,\xi^*} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}\xi_i + C\sum_{i=1}^{m}\xi_i^*$$
$$\text{s.t. } \omega^T\phi\left(x^{(i)}\right) + b - y^{(i)} \le \epsilon + \xi_i$$
$$y^{(i)} - \omega^T\phi\left(x^{(i)}\right) - b \le \epsilon + \xi_i^*$$
$$\xi_i, \xi^* \ge 0, i = 1, \dots, m.$$

Where $y^{(i)}$ is the training label.

- **Neural Network** of 3 fully connected layers with Relu activation function in the first two layers and linear activation function in the output layer
- **Gradient Boost** tree ensemble[2]:

Let F_0 be a constant model. $R\left(y^{(i)}, F_{m-1}\left(x^{(i)}\right)\right) = -\frac{\partial \text{Loss}\left(y^{(i)}, F_{m-1}\left(x^{(i)}\right)\right)}{\partial F_{m-1}(x^{(i)})}$. Train a regression tree sub-model h_m on R.

$$F_m(x) = F_{m-1}(x) + \alpha h_m(x)$$

Results

Training (39,980 examples) and validation (4,998 examples) splits were used to evaluate different models. Unused test split (4,998 examples) was used to provide unbiased estimate of error.

Model Name	R^2 Test/Train	MSE Test/Train
Linear Regression (baseline)	-5.1e13/0.690	2.4e13/0.148
Ridge Regression	0.660/0.677	0.161/0.155
Gradient Boost	0.586/0.712	0.196/0.138
K-Means + Ridge	0.675/0.699	0.154/0.144
SVR	0.690/0.777	0.147/0.107
Neural Network	0.669/0.725	0.157/0.132

Conclusion

- **Discussion:** The dataset contained too many features, which led to model overfitting, causing variation of error to rise. Feature importance analysis using Lasso regularization improved performance, and using more advanced models such as SVR and neural networks resulted in higher R^2 score for both validation and test sets. Given the heterogeneity of the dataset a 69% R^2 score for the best performing model (SVR) is a decent outcome.
- **Future:** The future work on this project can include (i) studying other feature selection schemas such as Random Forest feature importance, (ii) further experimentation with neural net architectures, and (iii) getting more training examples from other hospitality services such as VRBO to boost the performance of K-means clustering + Ridge Regression model specifically.

References

1. Inside Airbnb. (2018). Inside Airbnb. *Adding data to the debate.*. [online] Available at: http://insideairbnb.com/get-the-data.html [Accessed 7 Dec. 2018].
2. Cse.chalmers.se. (2018). [online] Available at: http://www.cse.chalmers.se/~richajo/dit865/files/gb_explainer.pdf [Accessed 7 Dec. 2018].
3. Csie.ntu.edu.tw. (2018). [online] Available at: https://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf?fbclid=IwAR3j-F3Un2p8avLZgKw6wHc2eyNQePAu7CzQA50uuWBkzTy840tjsjkLGBE [Accessed 7 Dec. 2018].
4. Textblob.readthedocs.io. (2018). TextBlob: Simplified Text Processing — TextBlob 0.15.2 documentation. [online] Available at: https://textblob.readthedocs.io/en/dev/index.html [Accessed 7 Dec. 2018].