House Pricing Prediction

Team

Shihao Tong 11318294
Zihang Cai 11319479
Danni Weng 11304053

MOTIVATION

Predicting housing prices is important because people in every community are facing difficulty in finding a home that they can afford. Our goal is to find a simple and also efficient model to predict the housing price using data from King County, CA, USA.

RELATED WORK

- Phan, T.D. [1] used neural network and polynomial regression to predict houssing price in Melbourne.
- Park & Bae used pure probability Naive Bayes model improved by Adaboost to predict housing price in Virginia.
- Wang et al. focus the use on purely geological data combined with neural network predicting the housing price in ShenZhen.

OUR SOLUTION

- Visulise 'location' and 'grade' as a heatmap and box plot respectively
- Polynomial regression with degree selection
- K-Nearest Neighbor
- Random Forest with XGBoost
- Model comparison in terms of RMSE(root of mean square error)

Data Visualisation

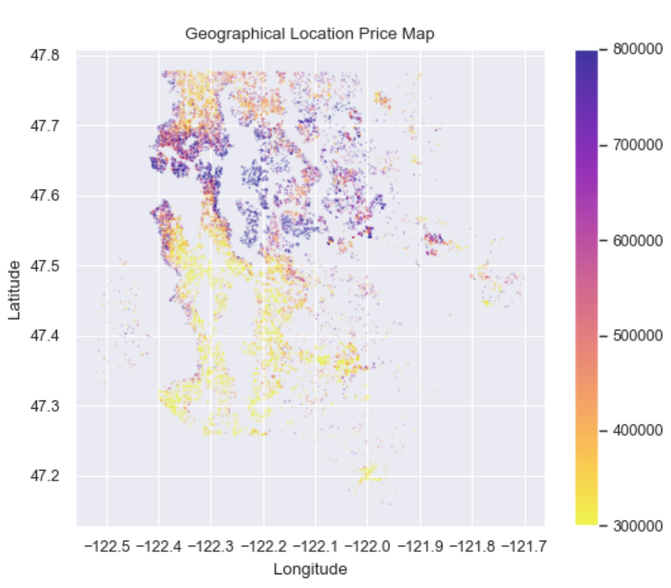

Figure 1: Geographical distribution of housing price in King County

The explanatory data analysis suggests two important pieces of information. Housing prices in King County show a clear pattern based on locations: The closer to lakes and sea, the higher the price. The second figure shows that the change in price distribution is more sensitive to the grade especially when grade is greater than 10.

#Grade`: Overall score for the quality of the renovations and building materials (score from 1 to 13 - a higher score indicates better quality materials used). This score is calculated according to the King's count scoring system.

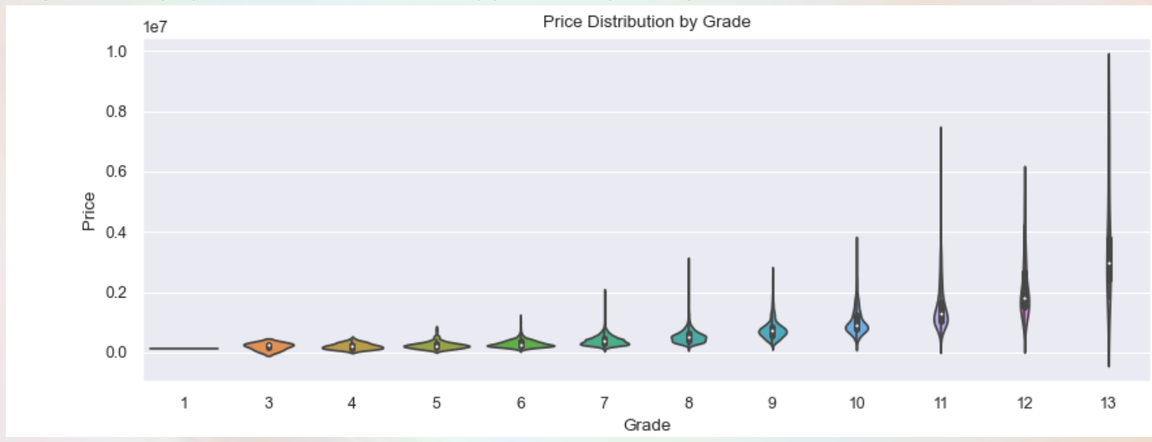

Figure 2: Housing Price v.s. Grade of Housing

Polynomial regression

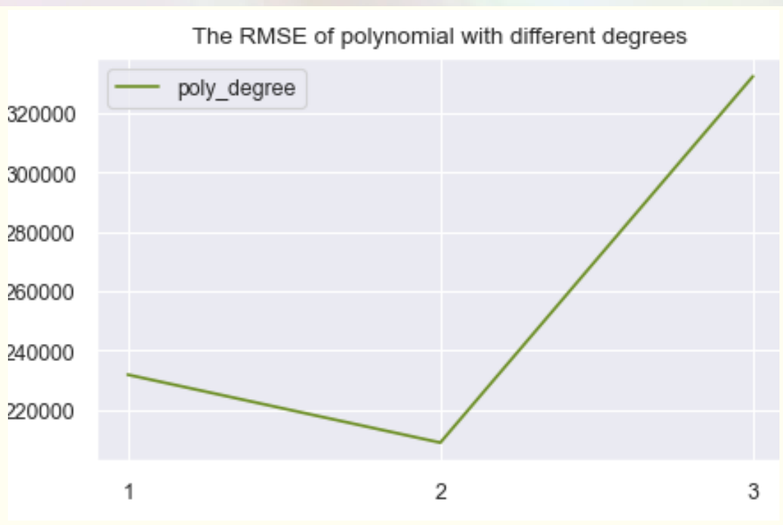

Figure 3: RMSE v.s. Highest Power of Polynomial Regression Model

Depends on our explanatory data analysis we found that many varibles do not have linear realtionship with price. Accroading to Phan T.D [1], polynomial regression is a valid approach. We tune our model and find that using King County's dataset, polynomial with degree 2 is the optimal model to fit. Degree of more than 3 are also tried but results in miserably big root mean square error (RMSE).

Random Forest with XGBoost

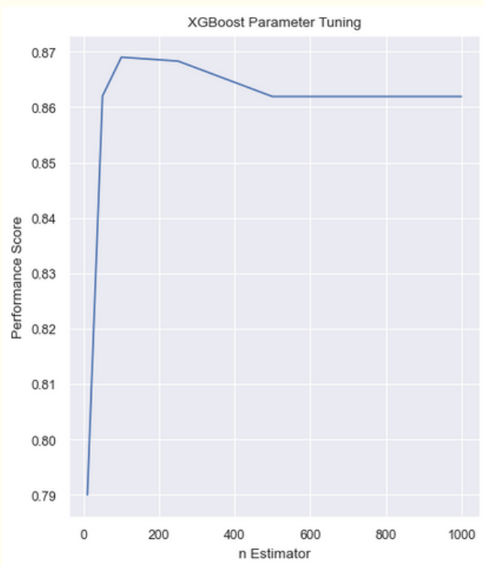

Figure 4: Performance Score v.s. n Estimator for XGBoost

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Boosting - It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. We will use XGBoost in this case.

n_estimator is the number of trees you want to build before taking the maximum voting or averages of predictions. The higher number of trees gives us better performance but makes the code slower. However, we found out that when we increase the n_estimator to 500, our model performance stopped improving.

K - Nearest Neighbor

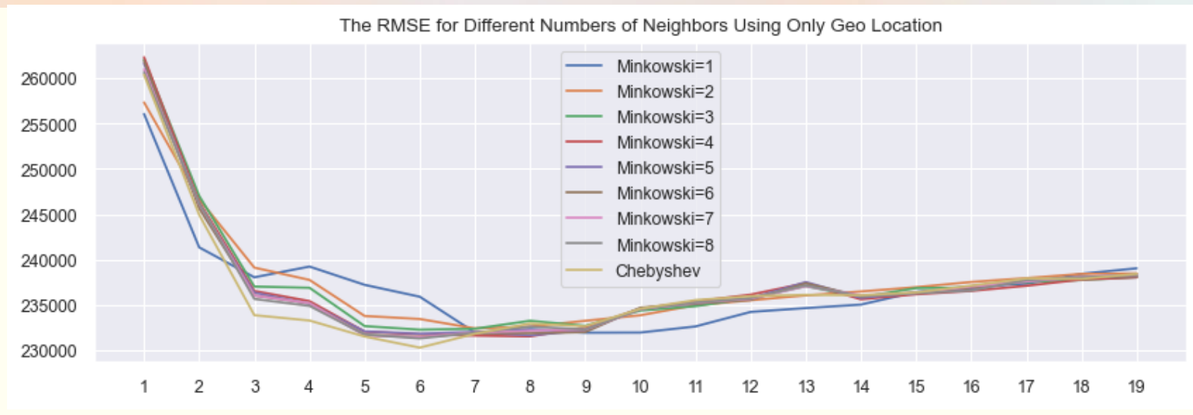

Figure 5: RMSE vs Difference number of Lags Using Longitute and Latitude Only

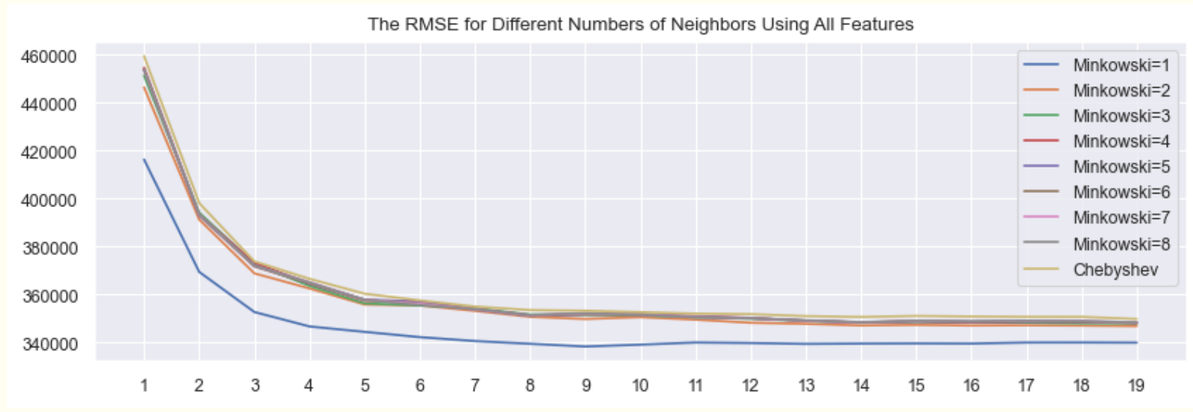

Figure 6: RMSE vs Difference number of Lags Using All Possible Features

Figure 3 and figure 4 shows the RMSE using different distance measure function : Minkowski distance from order 1 to 8 and the Chebyshev distance which are defined as

$$D(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

$$D_{\text{Chebyshev}}(x,y) := \max_i(|x_i - y_i|).$$

For purely geographical variables, the best K = 6, 7 or 8 for all distance measures and the Chebyshev distance function achieves the smallest RMSE at K = 6. For all variable cases, Minkowski distance beats all other distance measurement functions at any number of lags.

Comparison among Models

	RMSE
Polynomial regression	**208949.375**
Random Forest with XGBoost	**133128.411**
KNN (location)	**230262.632**
KNN (all)	**338329.645**

The precision of each model is evaluated by the root-mean-squared error (rmse) of the sale prices. Low RMSE is preferred. Our study shows that random forest XGBoost has the lowest RMSE among the 4 models. Also one thing worth notice is that for KNN model, adding more variables to distance function even harm the predicting accuracy. The KNN model should put more weights on the location variables instead tread all variables equally.

Future Works

The RMSEs are quite big in terms of our models, XGBoost performance is best among these weak prediction models. In order to produce a reliable prediction model, we would like to take full advantage of XGBoost, we are interested in exploring the hyperparameters tuning like booster, verbosity and nthread.

Reference

[1] Phan, T. D. (2018). Housing price prediction using machine learning algorithms: The case of Melbourne City, Australia. 2018 International Conference on Machine Learning and Data Engineering (ICMLDE). https://doi.org/10.1109/icmlde.2018.00017
[2] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040
[3] Wang, Z., Wang, Y., Wu, S., & Du, Z. (2022). House price valuation model based on geographically neural network weighted regression: The case study of Shenzhen, China. ISPRS International Journal of Geo-Information, 11(8), 450. https://doi.org/10.3390/ijgi11080450