Business Price Prediction in Google Local Review Dataset

Songlin Xu
UC San Diego
San Diego, USA, soxu@ucsd.edu

ABSTRACT

Business price in stores/restaurants on Google Map varies with different locations, working hours, and so on. People's review and rating could also reflect the price level of business on Google Map. By studying the correlation between business price and other factors like location and user subjective ratings, we could learn how the business price is determined all over the world.

In this project, we explore price prediction task on Google Local Reviews dataset[3, 8] according to five kinds of features: location, working hours, rating, review time, and visit time. We start with an exploratory analysis of the dataset to find potential correlations between business price and different properties as well as interesting findings. Then we explore machine learning models to tackle the price prediction task. Finally, we evaluate our models with ablation studies and describe related literature about this task.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Price Prediction, Geographic Information, Review Rating

ACM Reference Format:
Songlin Xu. 2023. Business Price Prediction in Google Local Review Dataset. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/xxxxxxx.xxxxxxx

1 DATASET

In this project, we aim to study the Google Local Reviews dataset [3, 8], which contains both geographic information and reviews about businesses from Google Maps. This large dataset includes 11,453,845 reviews, 4,567,431 users, and 3,116,785 businesses. Specifically, the metadata is composed of reviews, ratings, GPS coordinates, address, user information (places lived, jobs), timestamps, business category, business opening hours, etc. The original dataset is too large and some of them do not contain complete information such as GPS information or price. After filtering data samples with useless information (such as Nonetype), we split a new dataset that contains 52590 data samples for different places with various price levels. The overall dataset covers a time period of 14 years from 2000 to 2014, across different countries all over the world. The price level distribution is depicted in Fig. 1(2), where we could find that the price of business is divided into three levels: $ (level 1: 6.4% places), $$ (level 2: 60.3% places), and $$$ (level 3: 33.2% places). In this new dataset, we mainly consider five kinds of features related to business price: geographic location, working hours, rating, review time, and visit times, which are depicted in Table. 1.

Table 1: Feature Information

Feature	Format	Dimension	Range
Geographic Location	float	1×2	(-180,+180)
Working Hour	float	2×7	[0,24]
Rating	float	1×1	[1,5]
Review Time	integer	1×1	[2000,2014]
Visit Time	integer	1×1	[1,+inf)

- Geographic Location: composed of a 2-unit coordinates (longitude and latitude), representing a unique location on earth. Fig. 1(1) depicts the overall distribution of all GPS locations of business places in this dataset, where we could find that this large dataset covers different countries in different continents all over the world.
- Working Hours: composed of a 14-element list including open time and closed time of a business store/restaurant on Google Map in 7 days (1 week). Fig. 1(6) shows the distribution of overall working hours per week in different places in the dataset, where we could find that most places open for about 60-80 hours peer week.
- Rating: a float numeric value, which is the average of all ratings of one place on Google Map in the dataset. Fig. 1(3) demonstrates the distribution of user ratings in the whole dataset, where we could find that most users are willing to give high ratings (4 or 5 marks) to the places.
- Review Time: a float numeric value, which is the average year of all years of the reviews in one place. Fig. 1(5) shows the distribution of review years across the whole dataset, where we could find that most data are concentrated on the recent years (from 2010 to 2014), although the whole dataset includes reviews from 2000 to 2014.
- Visit Times: an integer, representing the number of times that users have left reviews in one place. This could represent the popularity of one place. Fig. 1(4) shows the distribution of visit times of different places in the dataset. where we could find that most places only have one visit in the dataset, although there are few places that have up to 22 visits.

1.1 Price Distribution with Geographic Information

Fig. 2(1) demonstrates the distribution of different business places all over the world. Each blue dot represents one place. Moreover, the

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/xxxxxxx.xxxxxxx

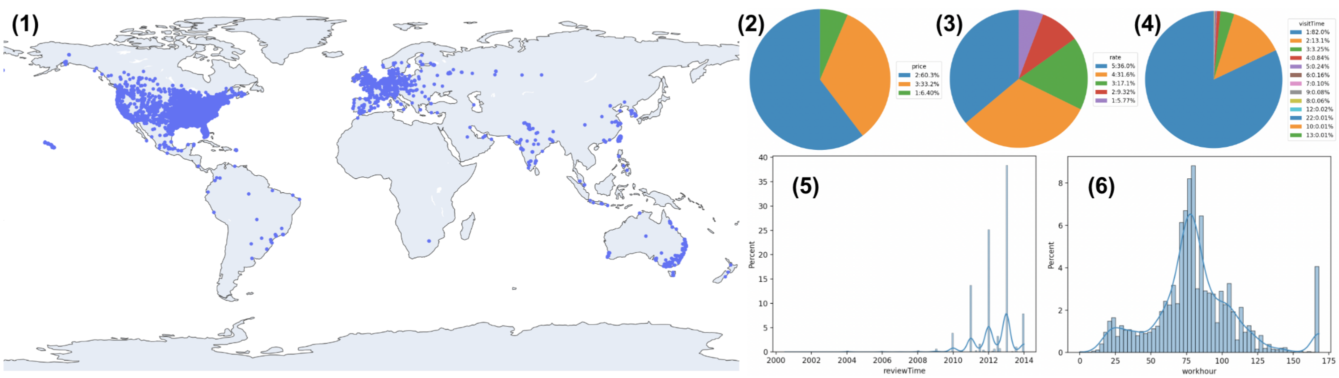

Figure 1: Basic statistics about the dataset, including distribution of GPS information (1), price level (2), user rating (3), user visit times (4), user review year (5), and place overall working hours in one week (6) in the whole dataset.

dot size stands for the price levels and larger dot represents higher price levels of the places. We could find different price levels in different geographic locations, which reveals potential correlation between geographic information and price levels. However, we then find that this correlation is different among different regions and it is hard to find a general rule to describe or represent this correlation. For instance, at first, we would like to cluster different locations according to the price levels. However, then we notice that even for the same state, the similar locations usually have different price levels. So it is hard to take a general manner to cluster locations. Therefore, here we directly leverage latitude and longitude as additional features in our price prediction task and our study results later show that this works well.

1.2 Correlation between Price and Working Hours

Fig. 2(2) shows the boxplot between price level and working hours per week, where we could find that there are significant differences in Price Level 1 v.s. Price Level 2 (P<0.0001), Price Level 1 v.s. Price Level 3 (P<0.0001), and Price Level 2 v.s. Price Level 3 (P<0.0001). The statistical significance is evaluated using post-hoc t tests. This reveals that working hour could be an important indicator/feature for price prediction since working hour is significantly different among places with different price levels. The potential reason behind the phenomena may be that different working hours may require different basic costs of business and will result in different price levels of the stores/restaurants.

1.3 Correlation between Price and Ratings

Fig. 2(3) shows the boxplot between price level and user average ratings in one place, where we could find that there are significant differences in Price Level 1 v.s. Price Level 3 (P<0.0001), and Price Level 2 v.s. Price Level 3 (P<0.0001). The statistical significance is evaluated using post-hoc t tests. This reveals that user ratings could be an important indicator/feature for price prediction since it is significantly different among places with different price levels. The potential reason behind the phenomena may be that different price may affect user ratings and user ratings may have an influence on the change of business price in turn.

1.4 Correlation between Price and Review Time

Fig. 2(4) shows the boxplot between price level and review years, where we could find that there are significant differences in Price Level 1 v.s. Price Level 2 (P<0.01), Price Level 1 v.s. Price Level 3 (P<0.0001), and Price Level 2 v.s. Price Level 3 (P<0.001). The statistical significance is evaluated using post-hoc t tests. This reveals that the review year could also be an important indicator/feature for price prediction since it is significantly different among places with different price levels. The potential reason behind the phenomena may be that business price is different due to the global financial situations in different years.

1.5 Correlation between Price and Visit Times

Fig. 2(5) shows the boxplot between price level and visit times, where we could find that there are significant differences in Price Level 1 v.s. Price Level 2 (P<0.001), Price Level 1 v.s. Price Level 3 (P<0.0001), and Price Level 2 v.s. Price Level 3 (P<0.01). The statistical significance is evaluated using post-hoc t tests. This reveals that visit times could be an important indicator/feature for price prediction since it is significantly different among places with different price levels. The potential reason behind the phenomena may be that the visit times could reflect the popularity of different places and different popularity is potentially correlated with the price levels of the business.

2 PREDICTIVE TASK

As depicted in our exploratory analysis, there are three price levels ($, $$, and $$$) in this dataset. Therefore, in this project, our predictive task is to classify the price level of one place, which should belong to one of the three price levels, given corresponding features as input. In order to evaluate our model in this predictive task, we will use different machine learning models for comparison, including Logistic Regression, SVM, Naive Bayes, Decision Tree, and Random Forest. We also utilize a dummy classifier to serve as the baseline model, which only performs classification according

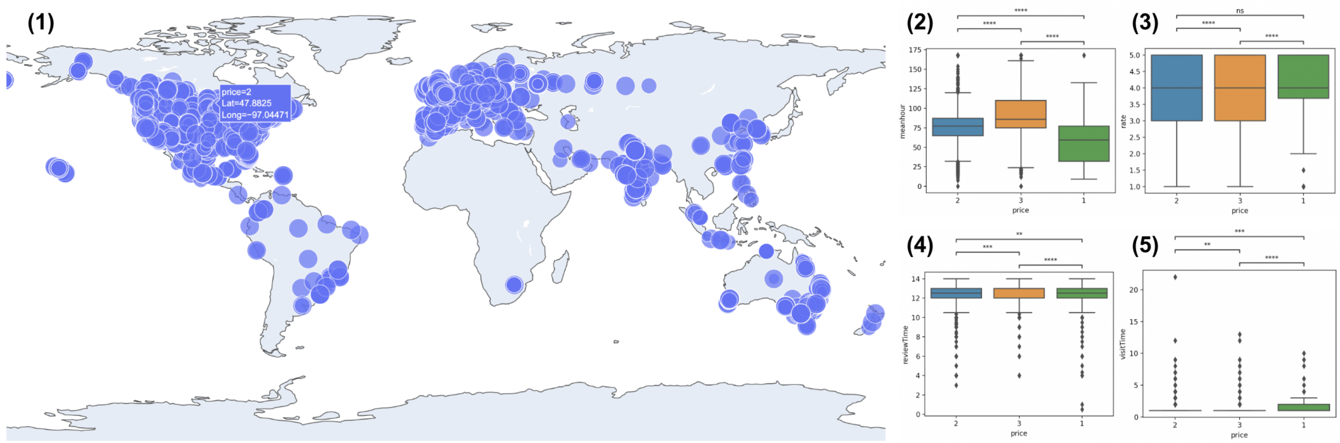

Figure 2: Correlation between business price level and geographic location (1), working hour per week (2), user average rating (3), user average review year (4), user average visit times (5). For (1), larger blue dot size represents higher price levels.

to the frequency of the features. We first shuffle and then separate the whole dataset into three parts: training set (50%), validation set (25%), and test set (25%). We use the validation set to find the optimal parameters of the models and then use test set to verify the effectiveness of our models in unknown dataset. For evaluation metrics, considering that this is a multi-class classification task, we use four metrics: accuracy, precision, recall, and F1-score to evaluate the feasibility and effectiveness of our model. Precision, recall and F1-score are calculated in an average-weighted manner.

For the features that we plan to use in this task, we decide to incorporate all the five features (geographic information, working hours, rating, review time, visit times), since our previous exploratory analysis has demonstrated that these features are statistically significantly different ($P<0.05$) among different price levels and could serve as important indicators of price levels. However, before feeding them into our models directly, we need to do preprocessing for the features. For working hours, we need to extract the open time and closed time each day peer week and get a $2 \times 7 = 1 \times 14$ dimension vector as the working hour feature vector. For user ratings and review years, since there might be more than one rating/review for the same place, we need to calculate the average ratings/review years to serve as the rating/review time feature. Similarly, for visit times, we count the number of visit times of each place to serve as the visit time feature. Note that for review time feature, we also subtract 2000 from the original review year to avoid the situation that the review time feature may introduce too much bias into the model. What's more, we also explore the effectiveness of data normalization using whether a Standard scaler or Min-Max scaler. However, our study results show that the normalization will lead to worse performance, which will be discussed later.

3 PREDICTION MODEL

We mainly use Random Forest model to solve the price prediction task. Random Forest is an ensemble learning model which will construct a multitude of decision trees during training according to the features and then use this trained decision trees to make prediction in unknown dataset. The reason why we decide to use this model is that it has good explanation ability and is not a black-box like neural network. In addition, this model does not require a large amount of data for training to achieve good results. It also has good performance to incorporate features with different specific information for prediction. The weights of Random Forest model is **optimized** during the model training process in the training set. Other hyperparameter settings follow the default settings in the RandomForestClassifier in scikit-learn[9]. The current dataset size is large enough to deal with the model training process and Random Forest is not a data-hungry model to acquire large amounts of data for gradient descent optimization. Therefore, we currently do not have specific problems for scalability and overfitting issues. However, we indeed have other issues due to noise and missing data. Some data samples do not contain price information or other features like review and ratings, which will be removed in our dataset. Moreover, for some bad points and noise data like GPS data which is beyond the scale of the earth, we also remove them directly in our dataset.

In addition to the Random Forest model, we also use other machine learning models for comparison, including Logistic Regression, SVM, Naive Bayes, and Decision Tree. Logistic regression model is one model used in our class. We use it to compare with the performance of the Random Forest model. We also use a dummy classifier to be the **baseline** model, which only consider the frequency of features for prediction. There are also strengths and weaknesses of the different models being compared. Logistic regression is very efficient for training and is easy to interpret, but it creates linear boundaries and has a strong assumption of linearity between input and output. SVM is effective in high dimensional spaces but is not suitable for large datasets. Naive Bayes does not need much training data but may not work well in the dataset with different data distribution. Decision trees could capture nonlinear relationships but is not stable even in a small change in the data. These models mentioned above are also from our class which are relevant to our predictive task. However, some other models in our class like latent-factor model and collaborative filtering may be inappropriate for this predictive task. The reason is that latent-factor model and collaborative filtering are good at user-item pair in recommendation systems. However, in our dataset, although we also have user ratings, we are not focusing on the user-item pair. In addition, for most samples, there are only one or two same user-place pair, which could not support the effectiveness of latent-factor model and collaborative filtering model.

There are also some unsuccessful attempts in our exploration. For example, at first, we directly fed the overall working hour per week into our model for prediction. But we found that the performance was no so good. Then we found that a more fine-grain working hour feature vector that incorporates both open time and closed time per day could provide more useful information for prediction and achieve better performance. In addition, we also tried to take this task as a regression problem for price prediction. But then we found that the error was large and a classification model was more suitable to tackle this task.

For model evaluation, as depicted in the previous section, we use the test set to evaluate the performance of price prediction model. Moreover, we harness accuracy, precision, recall, and F1-score as the evaluation metrics to evaluate the feasibility and effectiveness of our model in this multi-class classification task. Precision, recall and F1-score are frequently used in a binary classification task but we calculate them in an average-weighted manner to deal with the multi-class classification task. In addition, to demonstrate different feature importance and the role of different feature representations to identify which ones are more effective in the prediction task, we will also compare the prediction performance in four evaluation metrics using different types of feature vectors as input in our model. The results will be depicted in the result section.

4 LITERATURE

In this project, we use an existing dataset named Google Local Reviews, which originally aims to investigate using translation-based factorization machines for sequential recommendation[3, 8]. This Google Local Reviews dataset was used to evaluate the feasibility and effectiveness of the proposed translation-based factorization machines for sequential recommendation. In addition to Google Local Reviews dataset, there are also some other similar datasets in the existing work[1, 2, 4–7, 10]. For example, [5] utilized Yelp dataset to predict the helpfulness of online restaurant reviews using machine learning. [7] also presented a Virginia housing dataset for housing price prediction in Fairfax county. Similar work in housing price prediction task and datasets with machine learning also includes [6] (house price prediction for real estate), [4] (prediction on green building prices), [10] (Airbnb price prediction), [2] (location-centered house price prediction), and [1] (price prediction of peer-to-peer accommodation).

There are actually different state-of-the-art methods for different previous works. For example, in [4], the best method is Decision Tree. But for some other works, the optimal algorithm is neural network. The effectiveness of different models also varies with different datasets, features, and target problems. Even for the same price prediction task, different features and datasets may require different kinds of models to achieve state-of-the-art performance.

Table 2: Prediction Performance using Different Models

Model	Accuracy	Precision	Recall	F1-Score
Baseline (Dummy)	0.6181	0.3820	0.6181	0.4722
Logistic Regression	0.6572	0.6366	0.6572	0.6166
SVM	0.6480	0.6226	0.6480	0.5616
Naive Bayes	0.5977	0.6163	0.5977	0.6007
Decision Tree	0.7595	0.7603	0.7595	0.7598
Random Forest	**0.8136**	**0.8112**	**0.8136**	**0.8101**

Among all of these related work mentioned above, our model selection is actually inspired by this work [4], which listed and compared the performance of different machine learning models including SVM, Decision Tree, Random Forest, and so on for price prediction on green building prices. In addition, as for feature representations, our work is inspired by the location-centered price prediction work [2] to deal with location features. Moreover, features like ratings and reviews are borrowed from [5] that predicts online restaurant reviews and other related features like review years are inspired by [7] which predict house price.

In conclusion, the existing work demonstrates the effectiveness of harnessing different reasonable feature representations and specific machine learning models to deal with price prediction task. Compared with these work, our project is different in target task and specific feature representations. Most existing work focuses on house price prediction but our work is about Google business price prediction. In addition, our feature sets include novel representations for working hours, visit times and user subjective ratings. But the similar thing between our work and previous work is that we both consider geographic information for price prediction. Another different finding is that Random Forest model is the best one in our dataset but Decision Tree model is the optimal one in some other works like [4]. This may result from different feature variables and datasets. However, one similar conclusion from existing work and our project is that we both find the importance of geographic information in price prediction task.

5 RESULTS AND ANALYSIS

We evaluate our Random Forest model compared with different machine learning models and also explore the effectiveness of different features for the price prediction task.

5.1 Prediction Performance in Different Models

As depicted in the previous section, we compare the prediction performance of our Random Forest model with other machine learning models including Dummy Classifier (Baseline), Logistic Regression, SVM, Naive Bayes, and Decision Tree, whose results are shown in Table. 2. Obviously, Random Forest model works well and has significantly better performance in accuracy (81.36%), precision (81.12%), recall (81.36%), and F1-score (81.01%), compare with other models. The baseline model (Dummy Classifier) almost does not work and has the lowest F1-score (47.22%) and Naive Bayes has the lowest accuracy (59.77%). Logistic Regression could work and has similar performance with SVM (about 65% accuracy) but much lower than Random Forest. The Decision Tree model also works

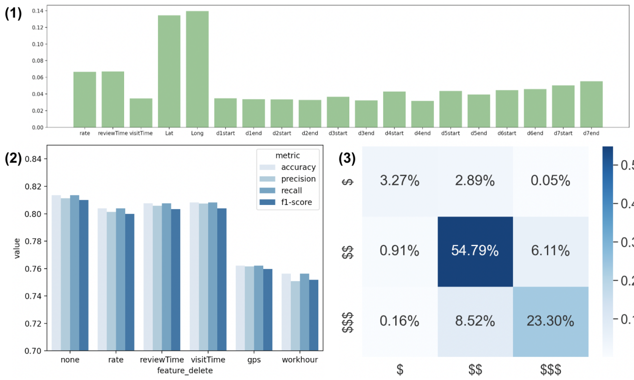

Figure 3: (1). Feature importance in price level prediction of Random Forest model. (2). Prediction performance of Random Forest model where different type of features are removed to demonstrate their effectiveness and importance. For example, the second group ("rate") of barplot represents the prediction performance where we remove the user ratings and use other features for prediction. The first group ("none") represents the prediction performance using all features for classification. (3). Confusion matrix of prediction performance using Random Forest model in test dataset.

well and has better performance (75.95% accuracy) than SVM but its performance is still worse than the Random Forest model.

The reason why the baseline model (Dummy Classifier) does not work (F1-Score lower than 50%) is that it only consider the frequency of features and could not capture complex relationships. For Logistic Regression model, it could work but it could only capture linear boundaries and therefore its performance is limited (about 65% accuracy). Similarly, SVM may not work well in the large dataset and could not capture complex and nonlinear relationships. Naive Bayes is too sensitive to different data distribution in the test dataset and therefore could not work well. Decision tree is easy to get biased due to a small amount of different data change and therefore the performance is lower than Random Forest model.

5.2 Interpretation and Feature Importance

In order to further **interpret** the parameters of the Random Forest model, we draw the feature importance plot in Fig. 3(1), which depicts the importance of different features as variables in the Random Forest model. Interestingly, we find that the geographic information (GPS location) has the highest importance in the prediction process of the Random Forest model. This phenomenon reveals the strong correlation between price level and business place locations. This is reasonable since different locations may result in different costs to support the business and therefore it may lead to different price. In addition, we also find that user rating and reivew time also play an importance role in price prediction model, which indicates that user feedback is also correlated with price level. This is reasonable as well since user subjective feelings may be affected by the price level and the price level may be affected by user ratings in turn as well. Among all working hours, we find that the working hours on weekends have larger importance that those on weekdays. This is another interesting phenomenon that reveals that the price level is more correlated with workings hours on weekends than those on weekdays. Probably because users would like to eat out more on weekends for relax after working for one week.

Table 3: Prediction Performance using Different Scalers

Scaler	Accuracy	Precision	Recall	F1-Score
None	**0.8136**	**0.8112**	**0.8136**	**0.8101**
Standard	0.7637	0.7663	0.7637	0.7514
Min-Max	0.6989	0.6970	0.6989	0.6731

5.3 Feature Ablation Study

We also run a feature ablation study to explore the effectiveness of different feature representations. Specifically, we compare the prediction performance of Random Forest model using different features as input variables. We first get the prediction results in four metrics (accuracy, precision, recall, F1-Score) with all five kinds of features (GPS, working hours, rating, review time, visit times) as input, which serves as the baseline performance. Then we delete one of the five kinds of features and use other four kinds of features as input and compare the prediction performance using the same model. We traverse the deletion of each of the five kinds of features and show the results in Fig. 3(2). Obviously, the baseline performance that does not remove any features have the best prediction results in four metrics. Then we notice that the prediction performance does not drop significantly when we remove the visit time feature, indicating that visit time is not an importance feature in price prediction model. When we remove rating or review time features, the prediction performance gets lower than before, demonstrating the importance of user subjective features. Moreover, when we remove GPS features, the prediction performance drops more significantly than before, which shows that GPS information works well in feature representations of this task and plays a very importance role. All of these results above are also consistent with the feature importance results in Fig. 3(1). Finally, prediction performance drops most significantly when we remove working hours features, revealing that working hours are the most importance kind of features. This conclusion is a bit different with the feature importance plot in Fig. 3(1). The reason is that the previous feature importance plot computes the importance of each element of the working hour features but our ablation study takes all working hour features as a whole.

5.4 Impact of Data Normalization

Moreover, we also explore the price prediction performance using different data normalization techniques, including standard normalization and min-max normalization. We use the Random Forest model with all five kinds of features as input under different data normalization techniques and show the results in Table. 3. Surprisingly, we find that data normalization even lead to dropped prediction performance. The reason may be that normalization will affect the weight of different features, which actually have different importance in the Random Forest model.

5.5 Final Configuration

Finally, the optimal configuration for price prediction is to use all five kinds of features as input using Random Forest model without data normalization, leading to 81.36% accuracy. We show the confusion matrix of predicted results in the test dataset in Fig. 3(3).

6 CONCLUSION

In this project, we explore different machine learning models with different feature representations and data normalization techniques for price prediction. There are still a few limitations. First, we do not find a better way to encode the GPS information, although current format could still work well in our model. A novel GPS location encoding method may achieve better prediction performance. In addition, the current dataset is unbalanced for price levels. But we do not take corresponding actions to deal with this issue, which could have achieved better prediction performance. We plan to tackle these problems in our future work to further push the limits of price prediction performance.

REFERENCES

[1] Muyun Dai, Wan Wang, and Lixin Miao. 2020. Price Prediction of Peer-to-Peer Accommodation Based on Sentiment Analysis and POI Feature Extraction. In *2020 The 11th International Conference on E-business, Management and Economics*. 15–19.

[2] Guangliang Gao, Zhifeng Bao, Jie Cao, A Kai Qin, and Timos Sellis. 2022. Location-centered house price prediction: A multi-task learning approach. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13, 2 (2022), 1–25.

[3] Ruining He, Wang-Cheng Kang, and Julian McAuley. 2017. Translation-based recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*. 161–169.

[4] Syafiqah Jamil, Thuraiya Mohd, Suraya Masrom, and Norbaya Ab Rahim. 2020. Machine learning price prediction on green building prices. In *2020 IEEE Symposium on Industrial Electronics & Applications (ISIEA)*. IEEE, 1–6.

[5] Yi Luo and Xiaowei Xu. 2019. Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of yelp. *Sustainability* 11, 19 (2019), 5254.

[6] Maryam Mubarak, Ali Tahir, Fizza Waqar, Ibraheem Haneef, Gavin McArdle, Michela Bertolotto, and Muhammad Tariq Saeed. 2022. A Map-Based Recommendation System and House Price Prediction Model for Real Estate. *ISPRS International Journal of Geo-Information* 11, 3 (2022), 178.

[7] Byeonghwa Park and Jae Kwon Bae. 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert systems with applications* 42, 6 (2015), 2928–2934.

[8] Rajiv Pasricha and Julian McAuley. 2018. Translation-based factorization machines for sequential recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 63–71.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[10] Siqi Yang. 2021. Learning-based airbnb price prediction model. In *2021 2nd International Conference on E-Commerce and Internet Technology (ECIT)*. IEEE, 283–288.