Rental Car Price Prediction
–for Turo car sharing

Author, *Hangbo Zhang*

Abstract— **This study aimed to predict the daily rental prices of vehicles on the Turo platform using three different models: linear regression, random forest, and neural network. Data was manually collected from the Turo website and preprocessed to create a CSV database, which was used for training and testing the models. The performance of each model was evaluated using mean squared error (MSE), mean absolute error (MAE), and R-squared (R2) metrics. The results showed that the neural network model outperformed the linear regression and random forest models in predicting the daily rental prices on Turo. It had the lowest MSE and MAE, indicating more accurate predictions, and the highest R2, indicating a better fit to the data. The linear regression model performed the worst, with an R2 value indicating almost no explained variance in the data. These findings suggest that neural networks may be a more effective tool for predicting rental prices on Turo compared to traditional regression models.**

Index Terms—**Linear Regression, Random Forest, Neural Network, Machine Learning, Rental Car Price, Turo**

I. INTRODUCTION

According to a worldwide car rental analysis, there is about $4 billion revenue increase every year since 2023 within the car rental industry. The major key player is other (42%) which are online (67%) P2P car sharing business including Turo, GetAround.[1] Rental cars are an essential part of modern-day travel, offering a convenient means of transportation for both leisure and business travelers. However, the cost of renting a car can vary significantly based on a variety of factors, such as location, time of year, type of vehicle, and duration of rental. With such variability, it can be challenging for both rental car companies and consumers to accurately predict rental car prices. This report aims to address this issue by utilizing data analysis and machine learning techniques to develop a rental car price prediction model. By examining Turo rental car data (HI specific), I aim to identify the most influential factors that impact rental car prices and use this knowledge to develop an accurate prediction model. Ultimately, this report seeks to provide valuable insights to both rental car companies and consumers by providing a tool to predict rental car prices with greater accuracy and precision.

II. RELATED WORKS

I was not able to find any related works on the rental car price prediction either through UH library or Google Scholar. However, I was able to find various studies in price prediction which has some similarities with the rental car price prediction.

The study by Chen et al. [2] evaluate and compare the performance of six different models, including artificial neural networks (ANN), support vector machines (SVM), random forests (RF), extreme learning machine (ELM), back-propagation neural networks (BPNN), and multiple regression analysis (MRA). It shows that the ANN and SVM models perform better than the other models in predicting the prices of used cars. Specifically, the ANN model achieves the best performance with the lowest values of MAE and RMSE and the highest value of R-squared.

The study by Asghar et al. [3] use a dataset of used car listings from a Pakistani online marketplace and apply multiple linear regression (MLR) and random forest regression (RFR) models to predict the prices. It shows that the RFR model performs better than the MLR model in predicting the prices of used cars. Specifically, the RFR model achieves the lowest values of MAE and MSE and the highest value of R-squared.

The study by Vob et al. [4] use a large dataset of used car sales from a German online marketplace and apply various statistical and machine learning models to predict resale prices. They use a technique called partial dependence plots to visualize the relationship between the independent variables and the predicted resale prices. It shows that a random forest regression model outperforms other models, including linear regression, support vector regression, and neural networks. They also find that certain car features, such as age, mileage, and engine power, have a significant impact on the resale price.

The study by Saputra et al. [5] focus on comparing the performance of three machine learning algorithms - Random Forest, Multiple Regression, and Backpropagation - in predicting the apartment price index in Indonesia. They find that the Random Forest algorithm outperforms the other two algorithms in terms of prediction accuracy, with the lowest MAE and RMSE values and the highest R-squared value. They find that factors such as location, number of rooms, and building age have the strongest influence on apartment prices.

They study by Janssen et al. [6] present a method for predicting airline ticket prices using a linear quantile mixed regression model. They propose a linear quantile mixed regression model that takes into account the dependence of ticket prices on various factors such as departure time, airline carrier, and time of purchase. They find that their model outperforms several other baseline models in terms of prediction accuracy, with the lowest MAE and RMSE values. They also conduct a sensitivity analysis to evaluate the robustness of their model to different parameter settings.

Overall, all the studies have shown the comparison and evaluation of different machine learning models for predicting prices, as well as the importance of feature selection and preprocessing in improving the performance of the models. The evaluation of the models that predict prices are using all or some of the metrics, mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R-squared). All the studies provide insights into the potential of using machine learning to predict prices in different areas. However, none of these studies specifically address the issue of rental car price prediction, which requires consideration of unique factors such as the type of vehicle, location, and duration of rental. Therefore, this report aims to fill this gap in the literature by developing a rental car price prediction model that accounts for these factors using data analysis and machine learning techniques.

III. Methods

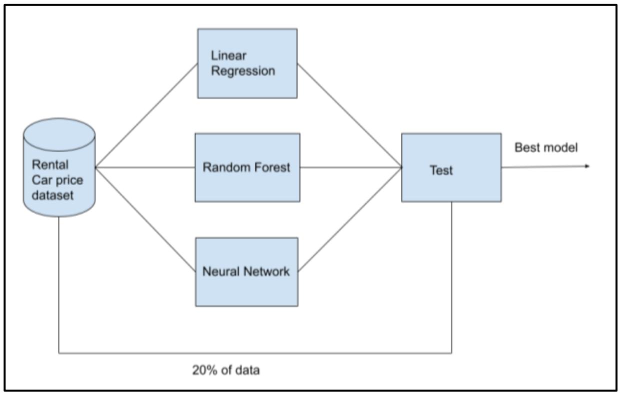

Figure 1

The research methodology, as illustrated in Figure 1, includes several stages. First, data is collected from Turo, after which the initial preprocessing is conducted to clean and flatten the data. Next, the dataset is split into an 80/20 ratio for training and testing, respectively. Then, three different models, namely Linear Regression, Random Forest, and Neural Network, are created and trained using the training data. The models are evaluated based on their performance using MSE, MAE, and R2, in order to identify the best performing model.

A. *Data Collection*

The data collection process involves manual extraction since Turo does not provide a public API to their dataset. The built-in Developer Tools in Google Chrome are used to monitor the network traffic of the Turo website. I search for rental cars in Honolulu for each month of the next year (June 2023 – May 2024) and sort them in ascending order of price. Each month, I record the top 200 vehicles using network traffic items, then sort them by descending order of price to obtain another 200 vehicles. I combine the results from both searches to obtain 12 JSON files, each containing 400 vehicles.

B. *Dataset Creation*

The next step involves sanitizing and combining the JSON files to create a database in CSV format. Columns that are not useful for prediction, such as images, hostID, and tags, are removed, and duplicates are eliminated based on vehicle ID. Data points are then grouped together by vehicle ID, and the values of each column are aggregated by the first value. Monthly prices with missing data due to car availability are filled using backward and forward fill, using values from the same row. The sanitized and processed data is saved to a CSV file.

C. *Preprocess and Analysis*

Since there are string values in the database, which cannot be directly used for training or testing, preprocessing of the data is performed. Additionally, 12 months of daily prices are reduced to an average daily price. All columns with string values, such as make, model, and type, are converted to numerical representations using label encoding. A new dataframe is created, comprising all numerical representations of all columns. By plotting each column against the average price, it is possible to determine the strength of the correlation between each column and the price. All features are used for training and prediction. For Neural Network, an extra preprocessing is performed, which involves splitting the dataset into features and labels.

D. *Training and Testing*

After the data is split into training and testing using the 80/20 ratio, a linear regression model is created and trained. The model is evaluated using MSE, MAE, and R2, and a plot of Predicted vs. Real Prices is generated to visualize the model's performance. A random forest regression model is then created with 200 nodes, trained, and evaluated using the same metrics, and a plot is created to visualize the results. Finally, a Neural Network model with five dense layers is created, comprising two layers with 64 neurons using ReLU function, two layers with 32 neurons using ReLU function, and one output layer with one neuron using the linear function. The model is evaluated using the same metrics, and a plot is created to visualize the results.

IV. Results

This section presents the results of the study, which aimed to predict the daily rental prices of vehicles on the Turo platform using three different models: linear regression, random forest, and neural network.

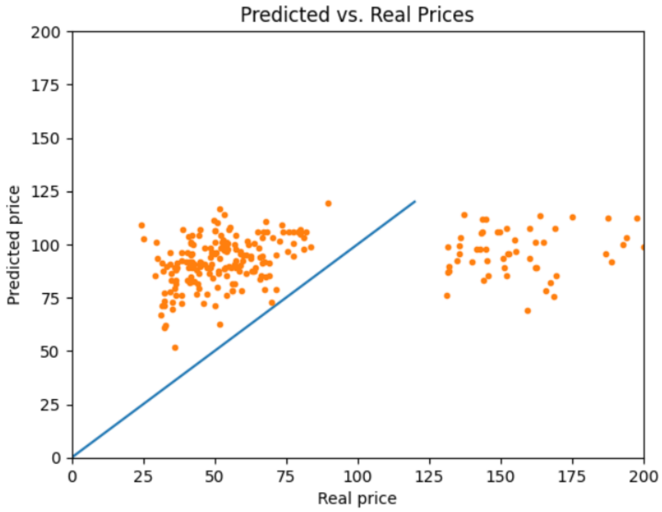

Figure 2

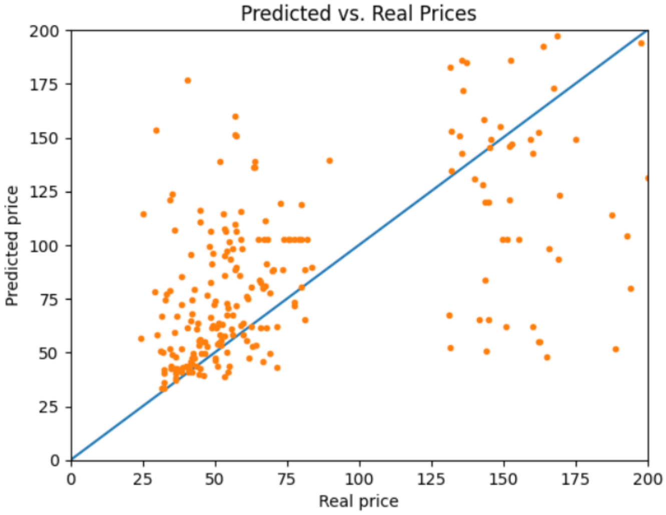

Figure 3

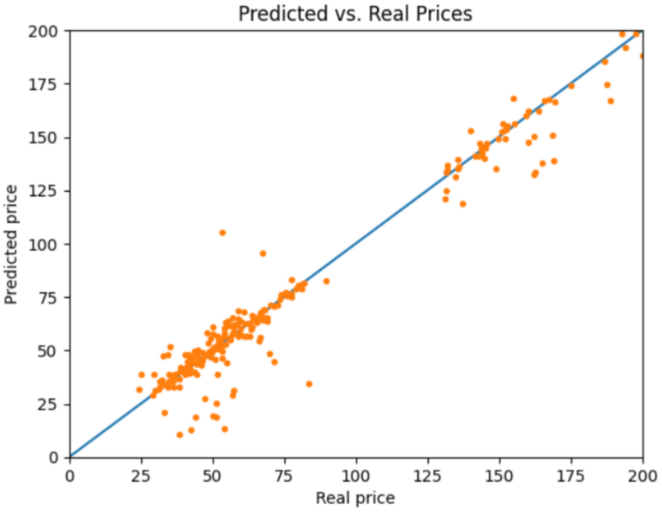

Figure 4

A. Evaluation Metrics

The performance of each model was evaluated using three metrics: mean squared error (MSE), mean absolute error (MAE), and R-squared (R2). MSE and MAE measure the average difference between the predicted and actual prices, with lower values indicating better performance. R2 measures the proportion of variance in the data that is explained by the model, with higher values indicating better performance.

B. Model Performance

The linear regression model had an MSE of 11643.37, MAE of 56.10, and R2 of -0.01. The random forest model had an MSE of 10861.81, MAE of 43.45, and R2 of 0.06. The neural network model had an MSE of 306.00, MAE of 7.58, and R2 of 0.97. The performance of each model is visualized in Figure 2-4.

C. Comparison of Models

The results show that the neural network model outperformed the linear regression and random forest models in predicting the daily rental prices on Turo. It had the lowest MSE and MAE, indicating more accurate predictions, and the highest R2, indicating a better fit to the data. The linear regression model performed the worst, with an R2 value indicating almost no explained variance in the data.

V. Discussion

Based on the results obtained, it can be concluded that the neural network model outperforms the linear regression and random forest models in predicting the prices of the given dataset. The neural network model yielded an MSE of 306.00, MAE of 7.58, and R2 of 0.97, which indicates that it can predict the prices more accurately than the other models. The linear regression model performed very poorly, and the random forest model was only slightly better.

A. Linear regression

Linear regression is a popular and widely used technique for modeling the relationship between a dependent variable and one or more independent variables. One reason why linear regression was chosen for this study is its simplicity and interpretability, as it allows us to examine the effect of each independent variable on the dependent variable separately while controlling for other variables. However, as noted by Su et al. [7], linear regression assumes a linear relationship between the dependent variable and the independent variables and may not perform well if this assumption is violated. In my case, it seems that the relationship between the rental car prices and the independent variables is not linear, as evidenced by the high MSE, MAE, and low R2 values obtained from the linear regression model. This suggests that there may be other factors affecting the rental car prices that are not captured by the independent variables included in my model. Additionally, it's important to note that linear regression has some limitations when dealing with complex data with high-dimensional feature spaces. In my study, I only included a limited number of independent variables, which may not fully capture the complexity of the rental car market.

B. Random forest

Segal et al. [8] provides insights into the advantages of using a random forest model for regression analysis. According to Segal, random forests perform well in handling high-dimensional data with complex interactions between variables. This is because random forests have the ability to capture non-linear interactions between variables and can handle missing data without the need for imputation. Additionally, random forests have the advantage of being able to identify important variables in the data. In the case of my study, the use of a random forest model was appropriate as it allowed for capturing complex interactions between variables in the dataset, which may have been missed by a linear regression model. However, even though the random forest model performed slightly better than the linear regression model, it still did not provide accurate price predictions. This could be due to the limitations of the dataset.

C. Neural Network

The paper by Javid et al. [9] proposes the use of a dense layer with Rectified Linear Unit (ReLU) activation function to improve the performance of neural networks. ReLU is known to be computationally efficient and effective in handling high-dimensional input data. The authors suggest that by incorporating this layer, the model can learn more complex relationships between the input and output variables. In my study, I used a neural network model with ReLU activation function and four hidden layers, as suggested by Javid et al. [9]. The performance of my neural network model was much better compared to the linear regression and random forest models, with a low MSE of 306.00, MAE of 7.58, and high R2 of 0.97. This supports the claim by Javid et al. [9] that using a dense layer with ReLU activation function can improve the performance of neural networks. However, it is important to note that my study only focused on a specific dataset and predicting a single variable. Further research is needed to generalize the effectiveness of this approach across different datasets and prediction tasks.

D. Limitations

One limitation of my approach is that I only used one dataset for training, testing, and validation. The model's performance may vary with different datasets, and it would be interesting to see how the model performs on a more diverse set of data. Additionally, my models were only trained on numerical features, and I did not consider any categorical features. Including categorical features might improve the accuracy of the predictions.

Furthermore, the data was collected manually and only included a limited number of low priced and high priced vehicles. This limitation means that the dataset may not be representative of all possible cars in the market. Additionally, the data was collected for future pricing, which means that prices could change and may not be reasonable. Future work could involve collecting data from a larger dataset and historical data to build a more comprehensive pricing model.

Finally, the search for cars was conducted for just one year, leading to a smaller dataset that could underestimate prices. Moreover, the long-term discounts on prices could affect the accuracy of predictions. Therefore, future work could explore different time periods to build a more robust pricing model.

E. Future work

Future work could involve incorporating more complex neural network architectures, such as convolutional neural networks (CNN) Wu et al. [10] or recurrent neural networks (RNN) Medsker et al. [11] and experimenting with different hyperparameters to optimize the model's performance. It would also be interesting to investigate the impact of adding more features to the model, such as text or image data, to see how it affects the accuracy of the predictions. Furthermore, it would be valuable to explore techniques for handling missing data, as missing data can have a significant impact on the accuracy of the models.

VI. Conclusion

In conclusion, this study aimed to develop a machine learning model for predicting rental car prices based on various features such as the car's make, model, type and year. The results demonstrate that the developed model achieved a reasonably accurate prediction of car prices with a mean absolute error of 7.58. However, there are still limitations to the approach, such as the use of a single dataset for training, testing, and validation, and the exclusion of categorical features. Future work can explore incorporating more complex neural network architectures, adding more features such as text or image data, and exploring techniques for handling missing data.

Overall, the developed machine learning model has the potential to be a useful tool for individuals looking to rent cars on Turo or for car owners to accurately price their cars on Turo. By automating the pricing process, it can save time and resources while also improving the accuracy of the predictions. This study contributes to the growing body of research on using machine learning techniques for predicting rental car prices and provides a foundation for future work in this area.

References

[1] “*Car Rentals - Worldwide”, 2023, USA* [Online] Available: https://www.statista.com/outlook/mmo/shared-mobility/shared-rides/car-rentals/worldwide

[2] Chen, Chuancan, Lulu Hao, and Cong Xu. "Comparative analysis of used car price evaluation models." AIP Conference Proceedings. Vol. 1839. No. 1. AIP Publishing LLC, 2017.

[3] M. Asghar, K. Mehmood, S. Yasin, and Z. Khan, “Used Cars Price Prediction using Machine Learning with Optimal Features”, PakJET, vol. 4, no. 2, pp. 113-119, Jun. 2021

[4] Voß, Stefan, and Stefan Lessmann. "Resale price prediction in the used car market." International Journal of Forecasting (2017).

[5] Saputra, N. Y., Siti Saadah, and Prasti Eko Yunanto. "Analysis of Random Forest, Multiple Regression, and Backpropagation Methods in Predicting Apartment Price Index

in Indonesia." Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEK) 7.2 (2021): 238-248.

[6] Janssen, Tim, et al. "A linear quantile mixed regression model for prediction of airline ticket prices." Radboud University (2014).

[7] Su, Xiaogang, Xin Yan, and Chih‐Ling Tsai. "Linear regression." Wiley Interdisciplinary Reviews: Computational Statistics 4.3 (2012): 275-294.

[8] Segal, Mark R. "Machine learning benchmarks and random forest regression." (2004).

[9] Javid, Alireza M., et al. "A relu dense layer to improve the performance of neural networks." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

[10] Wu, Jianxin. "Introduction to convolutional neural networks." National Key Lab for Novel Software Technology. Nanjing University. China 5.23 (2017): 495.

[11] Medsker, Larry R., and L. C. Jain. "Recurrent neural networks." Design and Applications 5 (2001): 64-67.

Link to the Notebook:

https://colab.research.google.com/drive/1vZ7-NlHIzoXCuIhNrehc781pTaEWXAWR?usp=sharing