Final Report - Forecast Used Car Price

1. Introduction

1.1 Motivation

In recent times, used car prices have skyrocketed due to chip shortage and supply chain disruption and production delay of new cars. The idea for this project is to create a platform that can price cars based on their characteristics, historical prices and macroeconomic factors. The goal is to be able to generate an accurate pricing model that will allow the user to compare offerings in the marketplace and determine if cars are priced fairly vs the model. This would point out arbitrage opportunities, allowing the user to capitalize and create a more transparent market.

1.2 Problem definition

This platform would provide an advantage to the initial users, because they will have a better gauge for the true value of the cars in the marketplace. However, as the platform's user base increases, that edge will disappear as the market becomes more efficient at price discovery. In the end, the consumers will end up benefiting as the bid-ask spread narrows.

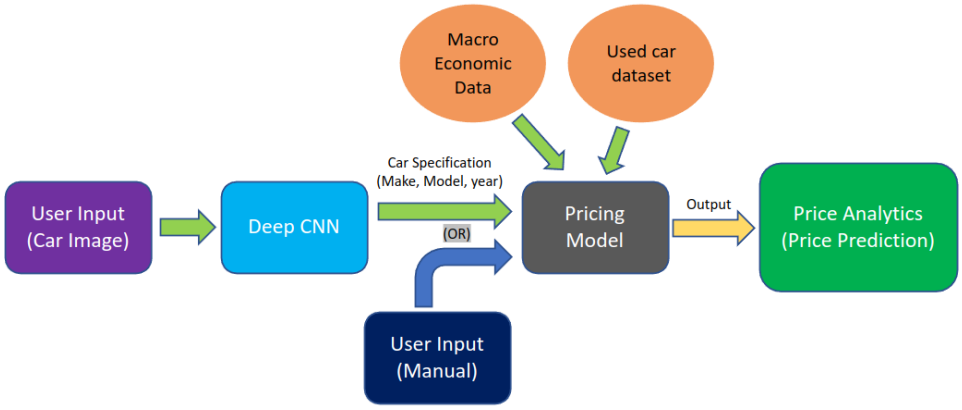
2. Survey

The most common approach for buying a used car is to go to an online platform, such as autotrader.com and search for cars based on the user's preferred parameters. Autotrader has listings for all sorts of vehicles, across the nation, so it is a significant improvement over how it used to be. People were limited to local dealerships and private parties for their used car selection. The internet opened the reach of sellers and buyers in the auto market, allowing for an increase in selection and competition. The issue with this approach is there is no information for the user besides what other users provide.

According to Erdem et. al (2009), the pricing model takes in different factors and varying degrees of impact on used cars price. It is suggested that the prices can be estimated accurately by implementing regression model as suggested by Kuiper et. al (2008). One of the most important features is the usage of a vehicle, particularly the mileage of a car. Thus, baking this factor dictates the value of the car over its lifespan (Engers et al., 2009).

We are also trying to use car image as input. The goal is to create 3d representation of images that may be used for categorization (Krause et al., 2013). This is helpful for our project because we are trying to add a similar feature where users can take pictures of their cars to get the best sell/buy price for their cars and also does analysis using the car image dataset that we intend to use in our project. CNN model will be used for image recognition (Zheng et al., 2017).

3. Proposed Method



3.1 Intuition

Comparing to used car valuation website like KBB, we creatively come up with a new method to analyze data and provide reasonable prices to the users. We also bring FRED data to take macro-economic factors into account.

- A new labelled used car image dataset (~2.3 million images) is created by leveraging the image links provided in the used car dataset.
- Image preprocessing is done and all the images are resized to 224 * 224 to use it with the VGG16 pretrained model.
- For the machine learning model, craigslist used car dataset and FRED data is combined based on the timeline.
- Preprocessing is done by eliminating the missing data and narrowing the number of attributes to seven.

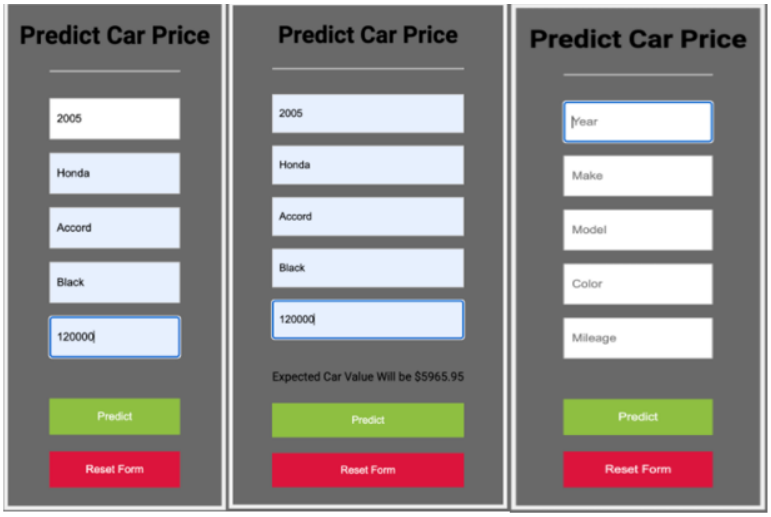
3.2 Approaches

We want to bring in economic data to enrich the model. In addition to car pricing and characteristics, we think variables such as inflation, interest rates, unemployment, GDP, etc. Will add value to the model. The last two years have shown that used car prices can fluctuate wildly because of economic factors.

The data will be obtained from FRED (Federal Reserve Bank of St Louis). Their API can be accessed from Python directly. They have a vast amount of data sets, some even going back to the early 1900's. We expect the robustness of this data to provide a significant amount of predictive power to the model. We hope that during volatile economic times, as we are experiencing now with high inflation, the model can match current prices in the market and be able to extrapolate into the future with accuracy, to allow users to make decisions on their personal situations and be better We are using Stanford cars dataset & VMMR dataset for the image data. The output of the deep convolutional neural network is provided as an input to the pricing model. A good Vehicle Make Model Recognition (VMMR) dataset is required to develop a reliable network that can provide correct Make, model and year of the car based on the input image provided by the user.

3.2 User Interface

For the user interface we made a HTML and CSS based frontend, that consisted of a form and a couple of buttons to submit the prediction and reset the form. We also made a backend application in python using the Flask library. Using Flask we created an API endpoint that would connect to a predict function that would load a pickled version of the model and read in the input form the data that was passed from the frontend to get a prediction. This value is then surfaced to the user on the front end as text beneath the form. The UI allows users to input their vehicle info such as the year, make, model, color and milage and expected car value displays at the bottom



4. List of Innovations

- Merging used car dataset and economic data
- Created a new labelled used car preprocessed image dataset.
- Leveraged the power of Linear regression and Random Forest models to train and get insights on the merged dataset.
- Fine-tuned pre trained VGG16 model on the newly created image dataset.
- Created a pipeline to input the results of the CNN model to the user interface to predict the car's value.

5. Experiments and Evaluation

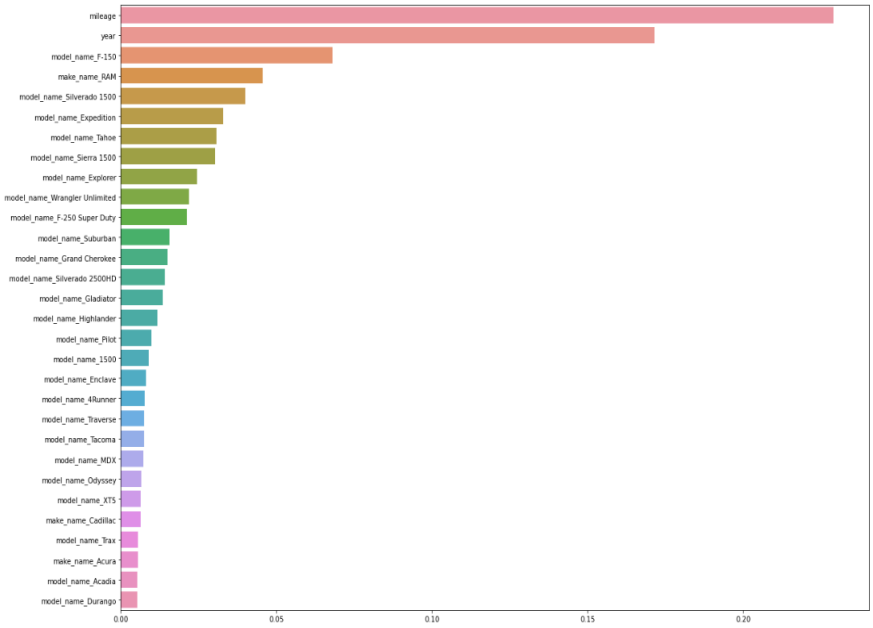
The main issue we encountered while working with the used car dataset was its size (10 gb), which made it difficult to run on our hardware. This issue made tasks more difficult because of the demand in computing power. For example, tasks such as splitting the data into training and test sets, cross validation, grid search, among others were either challenging or impossible to perform. Even with those difficulties,

we were able to run two models, linear regression and random forest. The random forest model performed better, with an RMSE of $3,173 and R^2 of 94.4%.

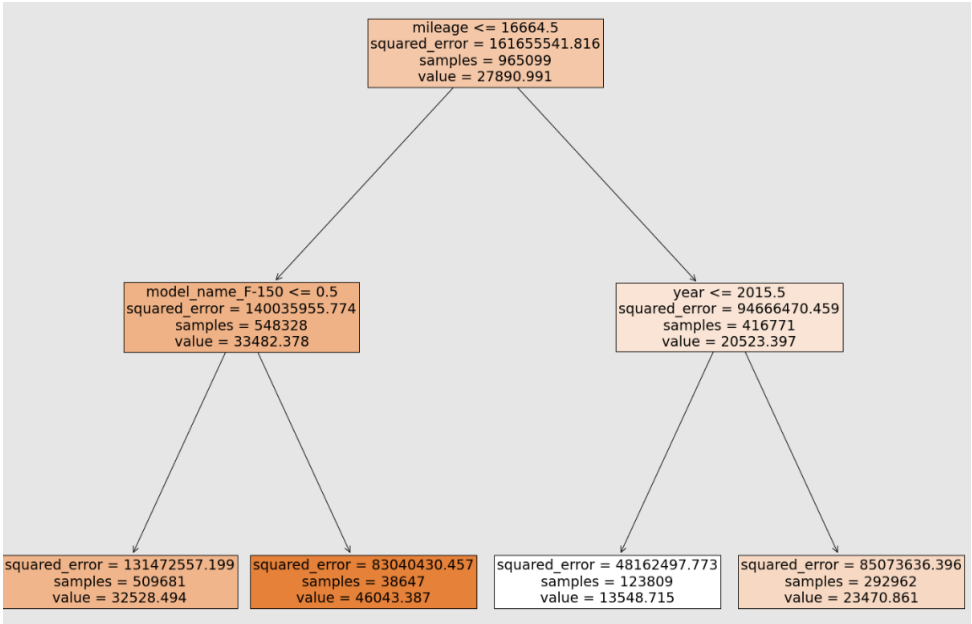
The next step in this experiment, now that we have a baseline model, is to perform feature selection to improve the metrics hopefully, but also to reduce dimensionality of the dataset and allow the hardware to handle the more demanding calculations required to fully validate the results. The plan is to eliminate features by utilizing VIF and stepwise forward selection, to achieve a balance between model performance and hardware limitations. We tried these approaches but ended up selecting the features for two reasons: based on the fields that are more commonly used in car shopping websites to make the UI user friendly and because of the predictive power these features showed when using the entire dataset. The features in the final model were make name, model name, listing color, mileage and year. The final model selection was random forest, with optimized hyperparameters, achieving a RMSE of $4,580 and R^2 of 85%. Even though the performance dropped, it is still acceptable and they allowed the model to use almost half of the original features. The reduced size of the model eased the limitations of the hardware and allowed us to do hyper parameter tuning, K-fold cross validation, feature importance, etc.

Regarding the economic variables we brought into the dataset, they turned out to have low predictive power. The main reason is the time frame of the listing dates. Even though the dates range from 11/1/2010 to 9/1/2020, the data is not evenly distributed. Most of the data points are in 2020, which means that the economic data integration was limited to a year. The dataset initially had over 3 million rows, so the 12 datapoints that matched the economic data caused a lot of repetition across rows. Additionally, the more extreme economic forces that moved used car prices happened in 2021, which is past the latest date in the dataset. For these reasons, the economic data did not add value, but if the data extended its distribution over a longer time period, we think it would have a stronger impact, especially factors such as inflation and income.

Important Features



Random Forest Tree



6. Conclusions and discussion

An interactive application for used car price forecast allows users to input either car image or conventional features. Users are able to see how each factor contributes to the historical price of a car and will be able to tell which of the current offerings provides the best value to purchase a vehicle. When it comes to the KBB model, we will know if we have improved on it if our model more accurately reflects real world prices, with the addition of economic factors.

Based on the bar graph of important features, we learned that mileage and year are the top 2 influencers on the price. Features below are mostly the names of the models. Top models are mostly pick-up trucks given the US auto market is heavily dominated by pick-up trucks. Due to the massive size of the decision tree, we were not able to show the full tree but the tree itself does a fantastic job of segmenting the inputs by different attributes. We did a couple of tests and found out the values were close to KBB price.

There are also many areas of improvement in this model. Firstly, we were not able to implement image processing algorithm into the model due to time constraints. The size of images was too big for our machines to accommodate and run through CNN model. With more time, we might be able to build more efficient code to process all images and produce desired output.

7. References

1. Hankar, M. (2022, June 22). Used car price prediction using Machine Learning: A case study. IEEE Xplore. Retrieved October 2, 2022, from https://ieeexplore.ieee.org/document/9800719

2. Kuiper, Shonda. “Introduction to Multiple Regression: How Much Is Your Car Worth?” Journal of Statistics Education, vol. 16, no. 3, 2008, https://doi.org/10.1080/10691898.2008.11889579.

3. 3D Object Representations for Fine-Grained Categorization Jonathan Krause, Michael Stark, Jia Deng, Li Fei-Fei 4th IEEE Workshop on 3D Representation and Recognition, at ICCV 2013 (3dRR-13). Sydney, Australia. Dec. 8, 2013. 4.

A Large and Diverse Dataset for Improved Vehicle Make and Model Recognition F. Tafazzoli, K. Nishiyama and H. Frigui In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops 2017.

5. Adrian Rosebrock, Deep Learning for Computer Vision with Python - ImageNet Bundle, PyImageSearch, Chapter 12: Case Study: Vehicle Identification, Pages 177:199, accessed on 5 October 2022

6. Venkatasubbu, P. (n.d.). Used Cars Price Prediction using Supervised Learning Techniques. From https://www.researchgate.net/

7. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248–255).

8. J. Fu, H. Zheng and T. Mei, "Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4476-4484, doi: 10.1109/CVPR.2017.476

9. H. Zheng, J. Fu, T. Mei and J. Luo, "Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5219-5227, doi: 10.1109/ICCV.2017.557.

10. Varshitha, J. (n.d.). Prediction of used car prices using artificial neural networks and machine learning. IEEE 2017 IEEE International Conference on Computer Vision (ICCV), 2022

11. Ceyhun Ozgur, Zachariah Hughes, Grace Rogers, Sufia Parveen, “Multiple Linear Regression Applications Automobile Pricing,” 2016 International Journal of Mathematics and Statistics Invention (IJMSI) E-ISSN: 2321 – 4767 P-ISSN: 2321 - 4759 www.ijmsi.org Volume 4 Issue 6 PP-01-10.

12. Xiling Wu, Caihua Zhang and Wei Du. ‘An Analysis on the Crisis of “Chips shortage” in Automobile Industry ——Based on the Double Influence of COVID-19 and Trade Friction,’ Journal of Physics: Conference Series, Volume 1971, 2021 3rd International Conference on Electronic Engineering and Informatics (EEI 2021) June 18-20, 2021 in Dali, China

13. Prieto, Marc, et al. “Using a Hedonic Price Model to Test Prospect Theory Assertions: The Asymmetrical and Nonlinear Effect of Reliability on Used Car Prices.” Journal of Retailing and Consumer Services, vol. 22, 2015, pp. 206–212., https://doi.org/10.1016/j.jretconser.2014.08.013.

14. Engers, M., Hartmann, M. and Stern, S. (2009), Annual miles drive used car prices. J. Appl. Econ., 24: 1-33. https://doi.org/10.1002/jae.1034

15. Erdem, Cumhur & Şentürk, İsmail. (2009). A Hedonic Analysis of Used Car Prices in Turkey. International Journal of Economic Perspectives. 3. 141-149