Forecast Used Car Price using Financial Data

Kavita Varma

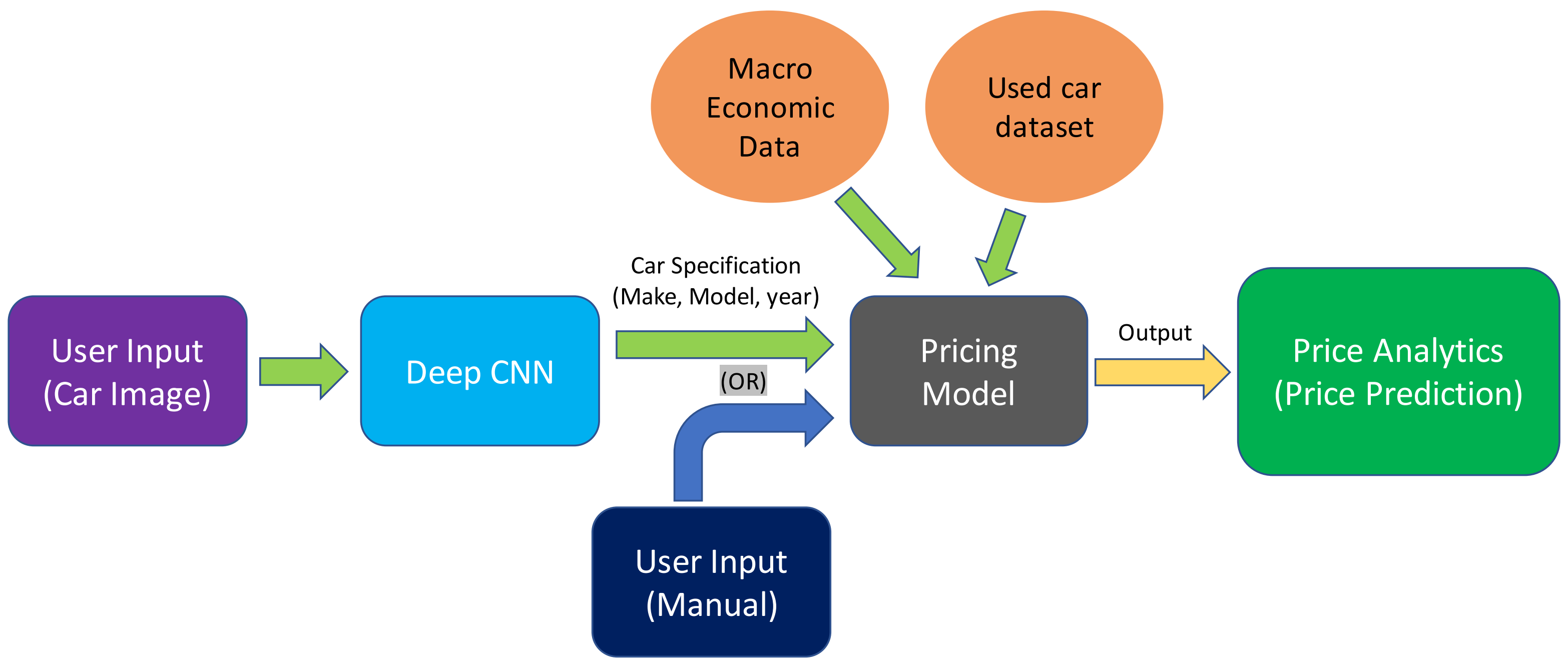
Motivation/Introduction

- Used car prices have skyrocketed due to chip shortage and supply chain disruption and production delay of new cars.
- The macro-economic factors like financial data can be leveraged to create a more accurate used car pricing model.
- The goal is to create a platform that predicts the car prices based on the characteristics and current economic factors.

Approach

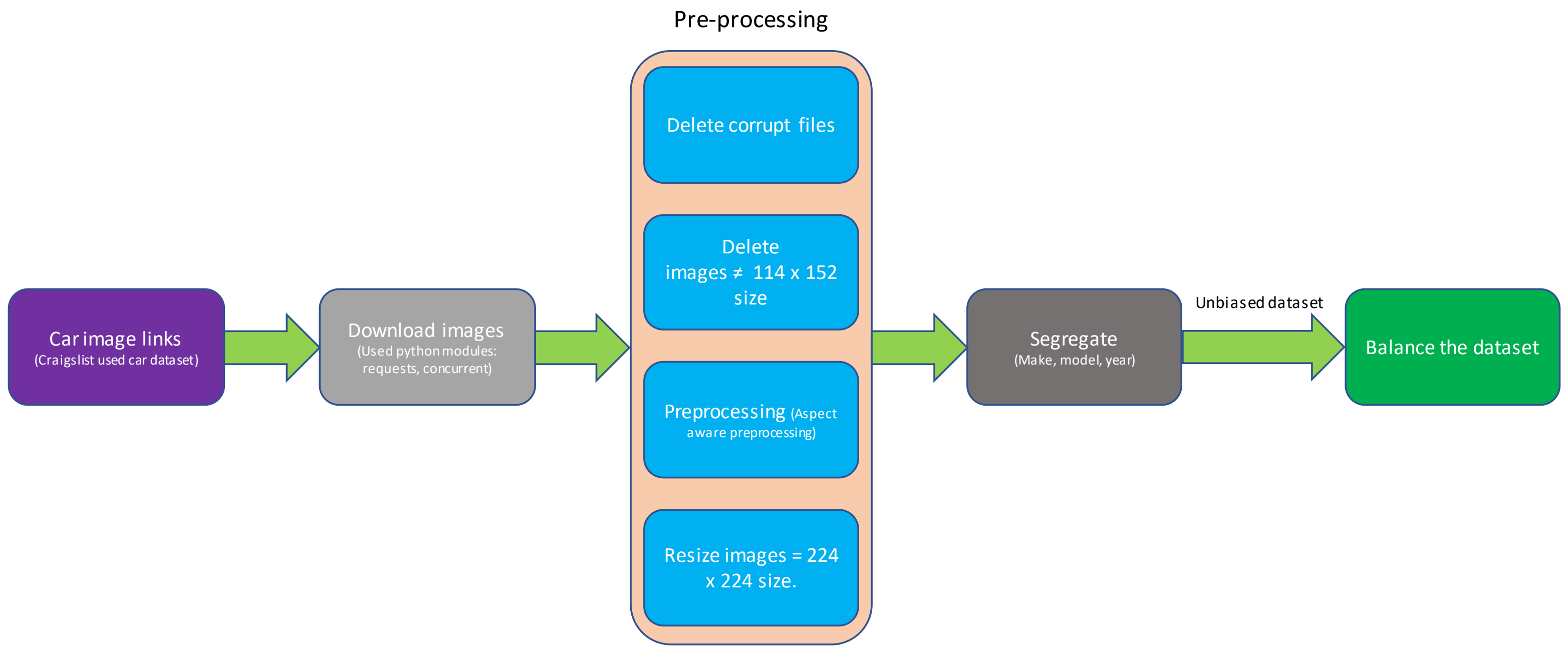
- In addition to car pricing and characteristics, I think variables such as inflation, interest rates, unemployment, GDP, etc. will add value to the model.
- The workflow is divided into two aspects:
  - **Deep learning module** and **Machine learning module**.
- Deep learning module is used to predict the car's **Make, Model and Year** based on the picture provided by the user.
- Machine learning module is used to predict the car's **value**.
- Preprocessed datasets are used for training a finetuned VGG16 model and a Linear regression model.
- User interface is created using HTML & CSS frontend and a Flask backend to interact with the trained linear regression model to predict the price.

Workflow



Data Collection / Preparation

- Datasets used for this project: 1) **Craigslist used car dataset** and **FRED (Federal Reserve Bank of St Louis)**.
- 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.


Pre-processing
Car image links (Craigslist used car dataset)
Download images (Used python modules: requests, concurrent)
Delete corrupt files
Delete images ≠ 114 x 152 size
Preprocessing (Aspect aware preprocessing)
Resize images = 224 x 224 size.
Segregate (Make, model, year)
Unbiased dataset
Balance the dataset

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.

Experiments and Results

Setup:

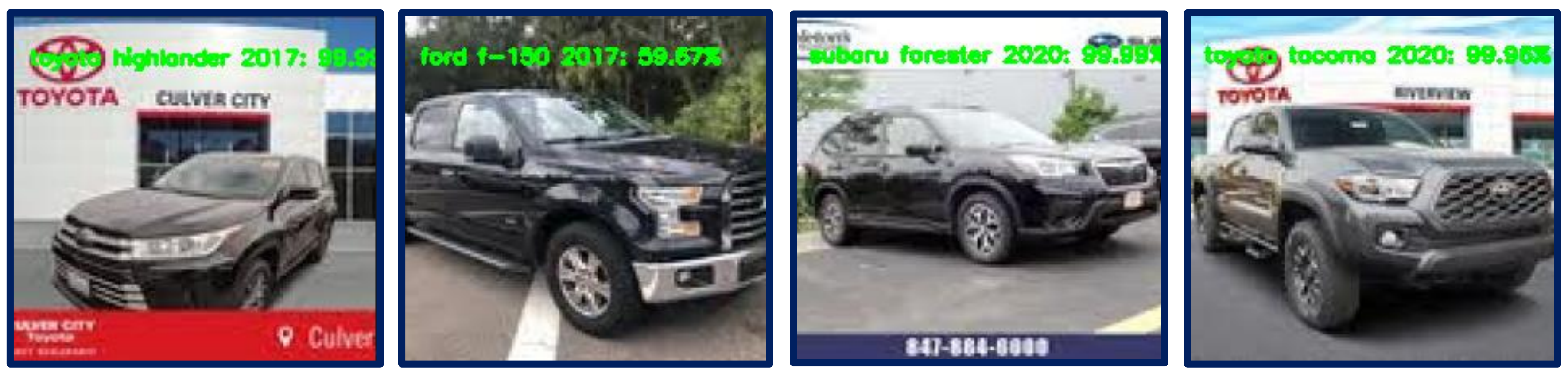
- Amazon's mxnet module with pretrained VGG16 weights for training the deep learning model.
- Jupyter python notebook with scikit-learn, pandas package to train the machine learning model.

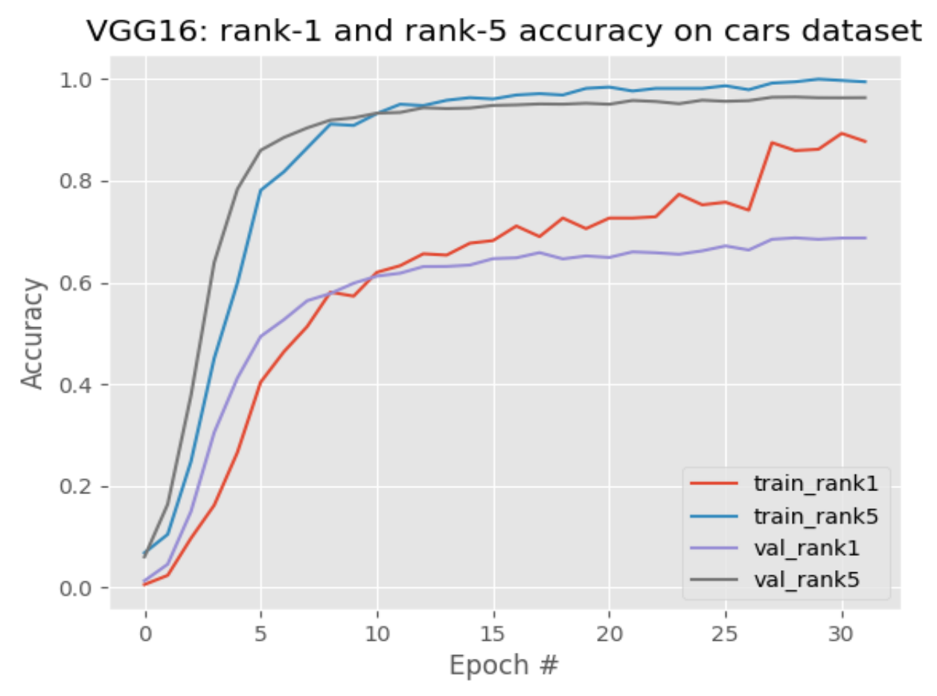
Results – DCNN model:

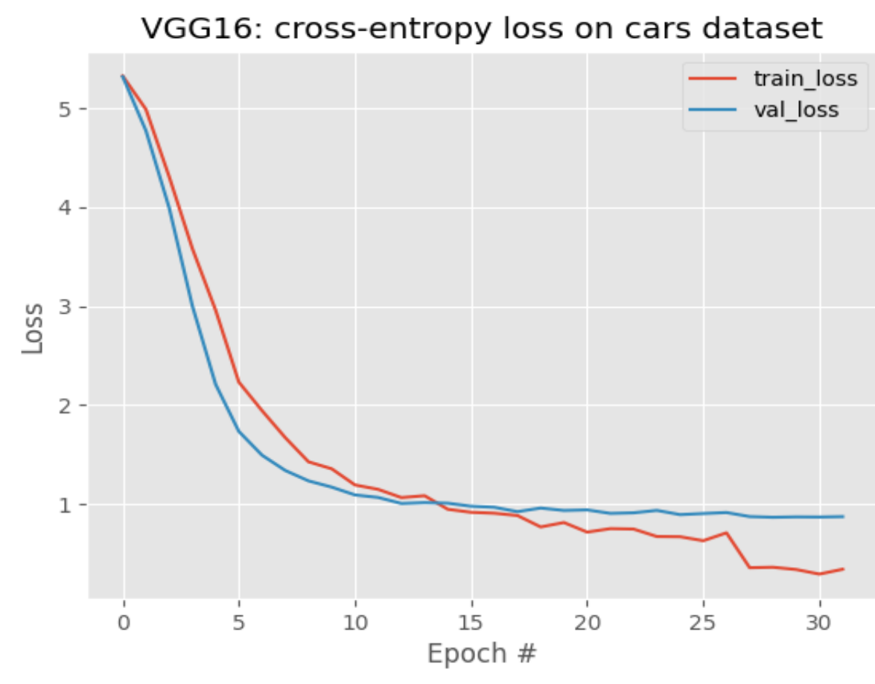
- Top-1 training accuracy of the finetuned VGG16 model is 88.5%
- Top-5 training accuracy of the finetuned VGG16 model is 99.2%
- Top-1 validation accuracy of the finetuned VGG16 model is 69.05%
- Top-5 validation accuracy of the finetuned VGG16 model is 96.5%

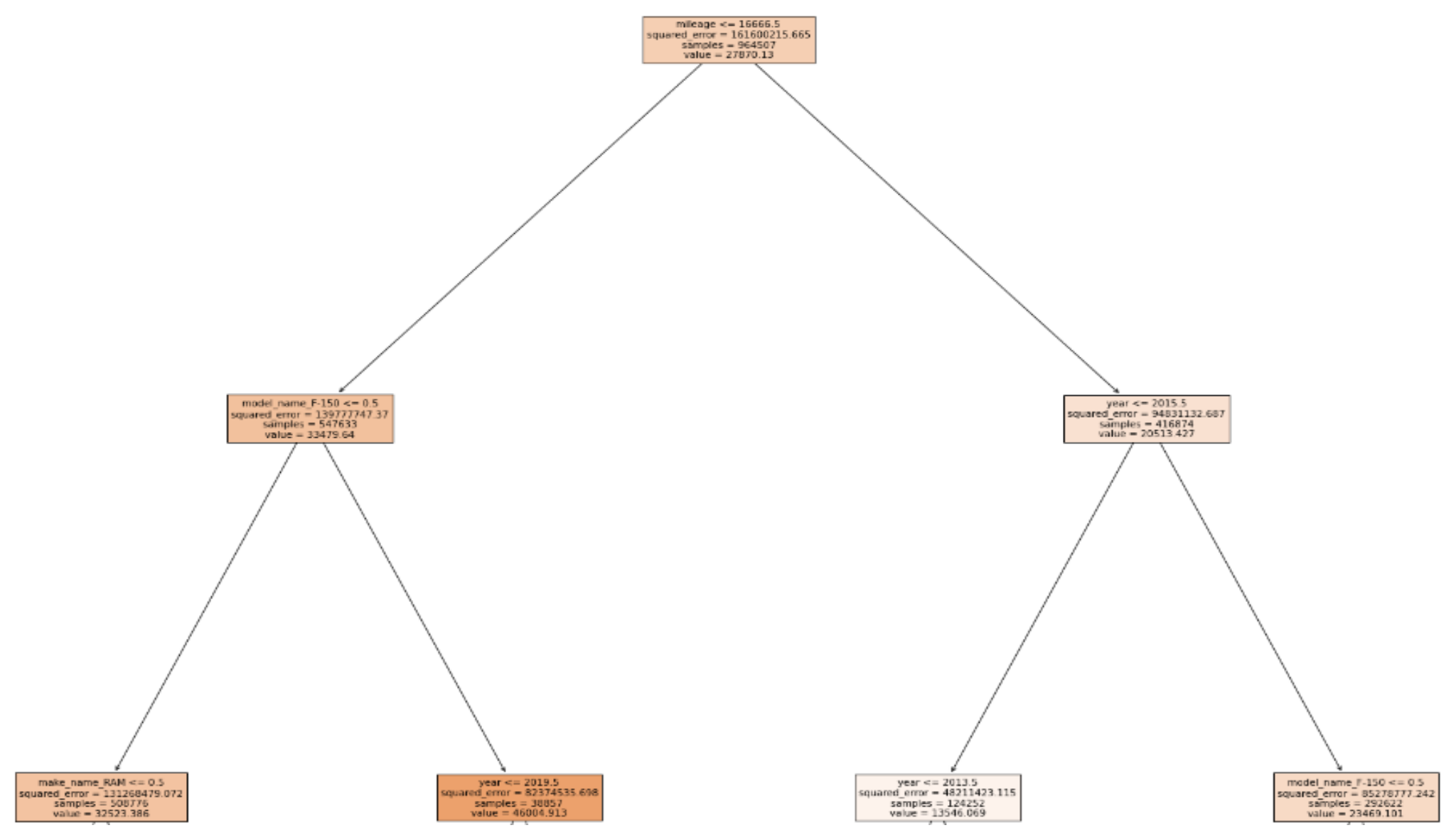
Results – ML model:

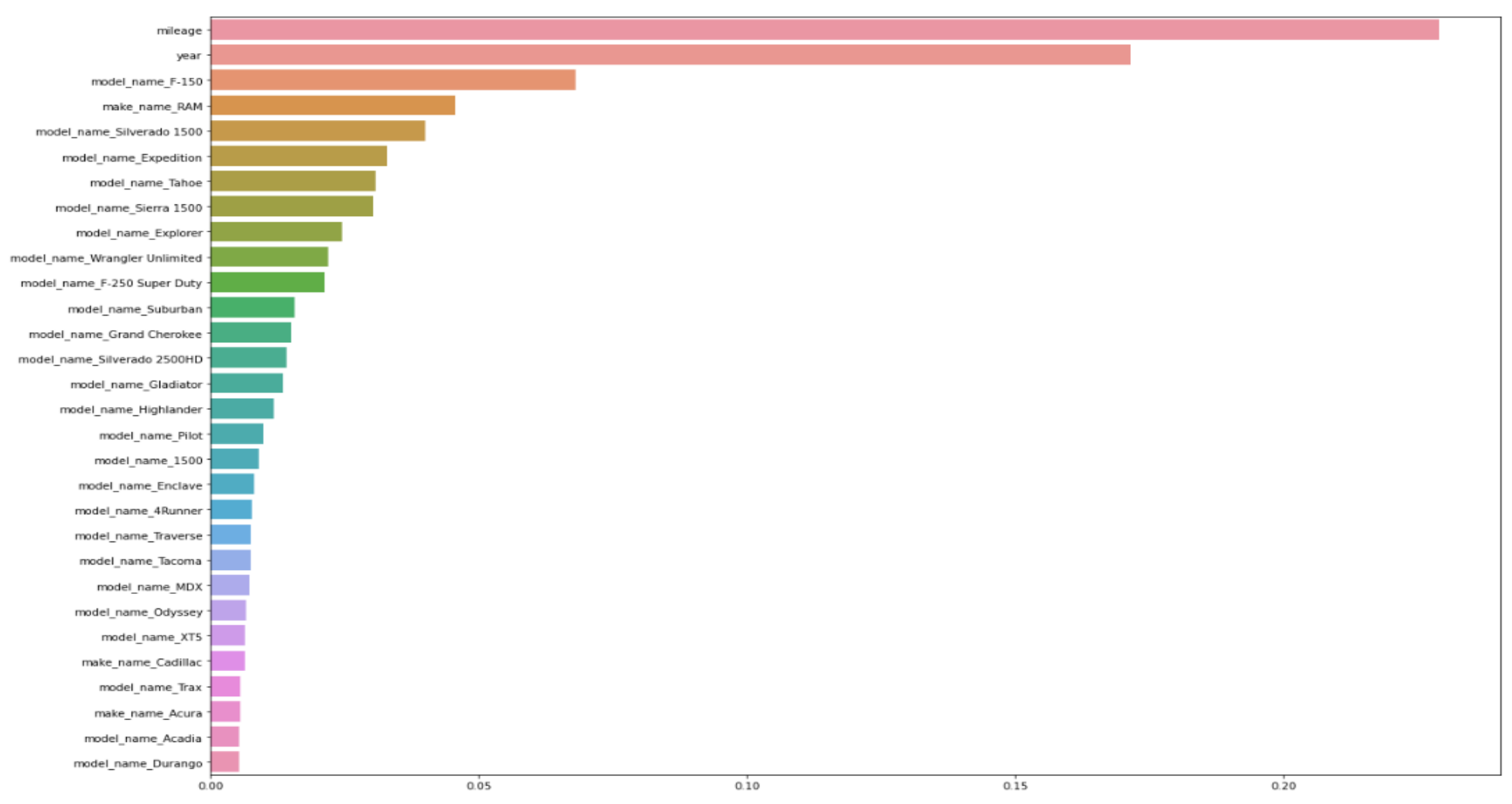
- Random forest performed satisfactorily, with an RMSE of $4,580 and R^2 of 85%.
- Features were selected based on their predictive power and their usage in car shopping web sites. Almost half of the features were dropped with this approach.
- Based on the bar graph of important features, we learned that mileage and year are the top 2 influencers on the price. The F-150 was 3rd, which is the most sold vehicle.
- Due to the massive size of the decision trees, we were not able to show the full tree
- Random forest did a fantastic job of segmenting the inputs by different attributes.
- This model's price prediction is comparable to KBB's evaluation

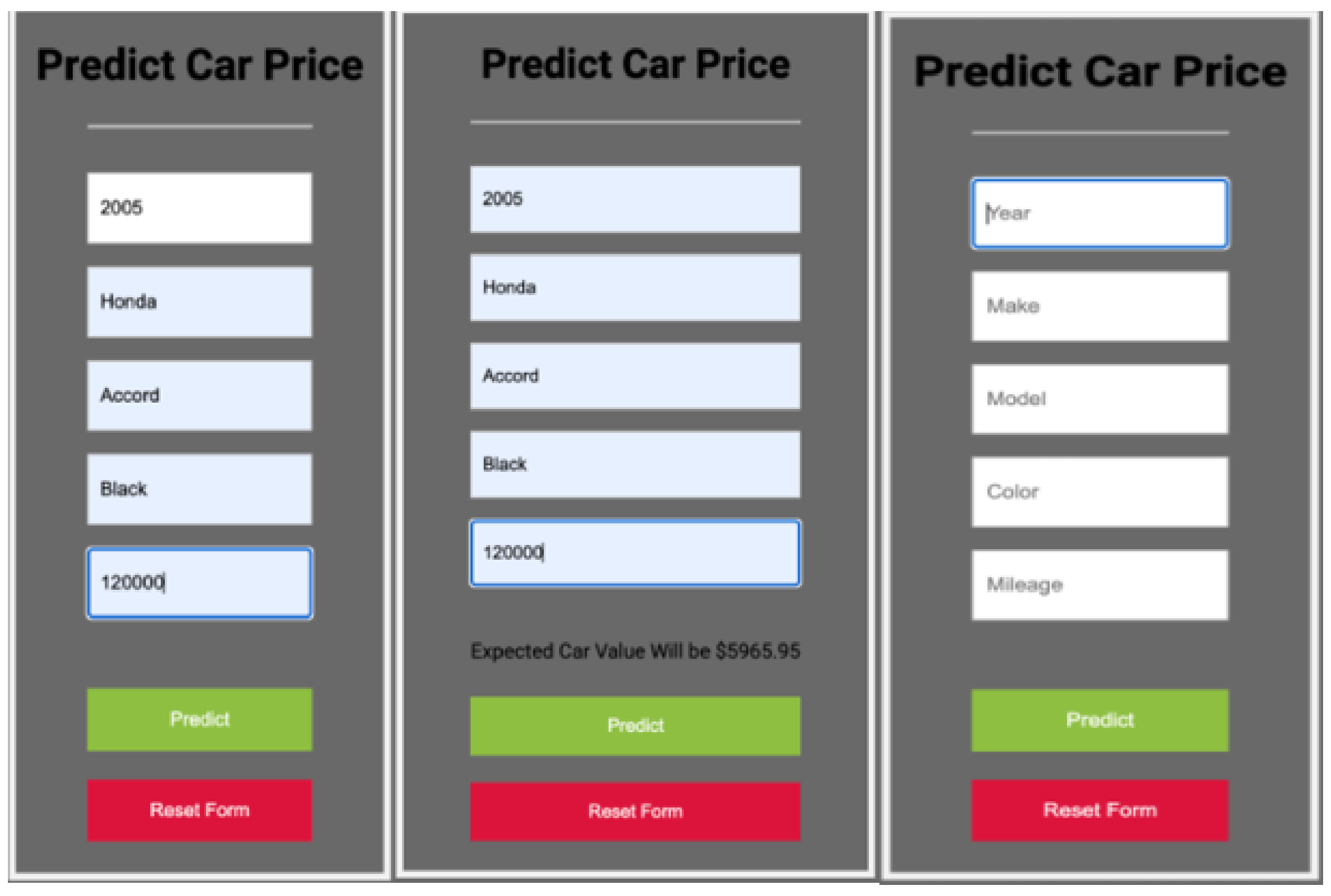

toyota highlander 2017: 99.9
ford f-150 2017: 59.67%
subaru forester 2020: 99.99%
toyota tacoma 2020: 99.96%


VGG16: rank-1 and rank-5 accuracy on cars dataset


VGG16: cross-entropy loss on cars dataset







Predict Car Price

2005

Honda

Accord

Black

120000

Predict Car Price

2005

Honda

Accord

Black

120000

Expected Car Value Will be $5965.95

Predict Car Price

Conclusion and Future Work

- The economic data imported into the dataset did not have predictive power, for two reasons: Even though the time frame for the listing date goes from 11/2010-9/2020, most of the listings fall in 2020. Given the dataset initially had over 3 million rows, a year worth of economic data caused a lot of repetition, which doesn't add explanatory power. Additionally, the extreme inflation data we've seen in recent years happened after the latest listing date, so that wasn't captured in the model.
- Finetuned VGG16 achieved a 96.5% top-5 accuracy. A complete pipeline with integrated deep learning model and the machine learning model would be next step.

Future work:

- Incorporating additional economic factors like inflation, any supply chain metrics to improve the accuracy of the pricing model
- An interactive application for used car price forecast which allows the users to input either car image or enter features and the data is fed into the machine learning pipeline to output and display the factors contributing to the historical price of a car and also provide an insight into the current offerings for best value.