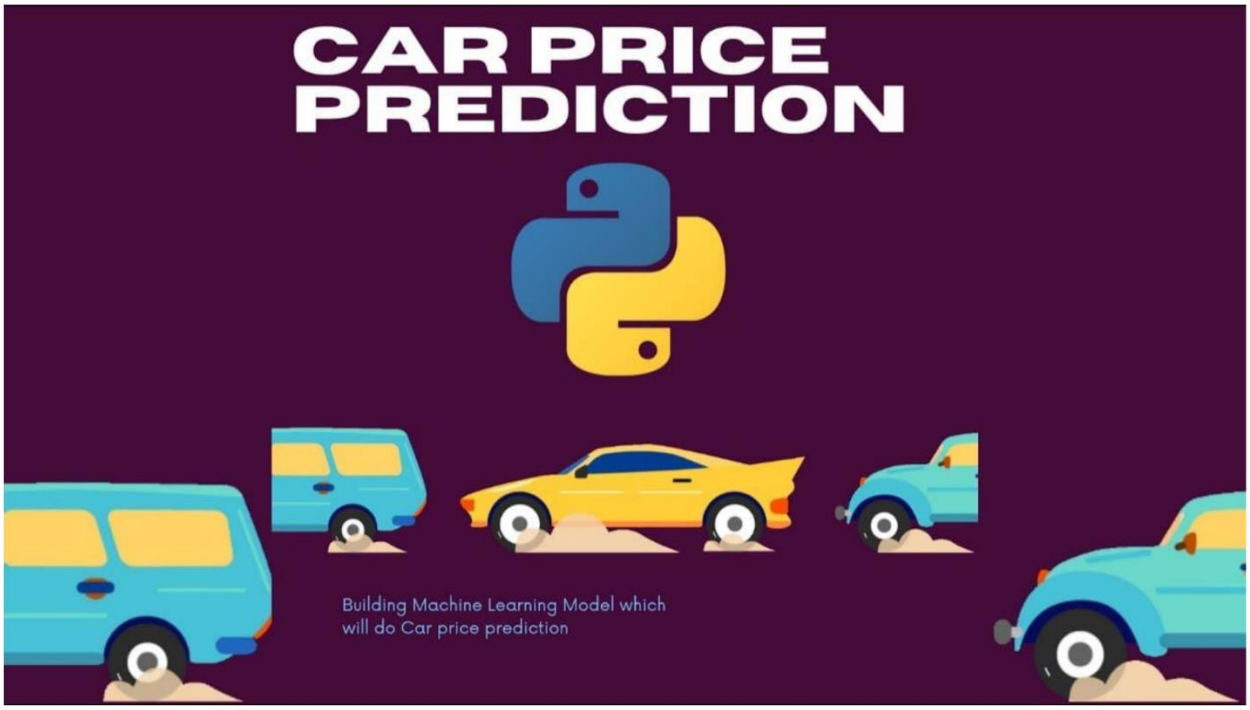
Project Final Report on Used car Price Prediction

Akhil Venu Gopal, Sasidev Mahendran



Data Preprocessing

Data Collection

For the initial phase of the project, we scraped data from www.cars.com , and created a dataset containing 25000 records. We used the python library BeautifulSoup for scrapping the data. We tried scrapping more detailed data features including color, drive gallon, mile per gallon, fuel type, safety, interior, convenience from the same website. Though we were able to do it, scrapping all of the above data for each of the 25000 records took a lot of time. So as an alternative, we had to take an open dataset on the same car price prediction (link). This is an historic dataset which contains most of the information about the car sales and is taken from Craiglist, which has the world's largest collection of used vehicle sales information. This data was also scrapped and converted into a .csv file from Craiglist and kept open for everyone on kaggle. There is no changes in the dataset and we are using the same for this phase as well. We are attaching the code for the data scrapping which we attempted in the first and second phase along with this report for reference.

Data Management

The data is downloaded in a csv format. This has been read into python using a pandas dataframe. The csv file had around twenty-six columns originally, containing columns including features, price, model, make, year, published year, url. Most of the columns in the dataset are features which can impact the price of the cars over time so we had to perform certain data cleaning and exploratory data analysis. The dataset consist of Id column which has unique id's for each record, url column which stores the corresponding craiglist url, region column saving the region of availability, region url which stored the craiglist url based on region, price column having the expected price, year column having the made year of the car, manufacturer column storing the name of manufacturer, model column having the model of the car, condition column having the condition of the car, cylinders column having the number of cylinders in the car, fuel column having the type of fuel used, odometer column having the odometer reading of the car, title_status column giving idea on the status of the record, transmission column storing the cars as manual/automatic, VIN column saving the engine details which is unique for each car, drive

column having the type of the drive like four wheel drive, front wheel drive or back wheel drive, size column saved with the size category of the car, type column having the type of car (truck, SUV, Sedan), paint_color column which have the exterior color of the car for sale, image_url column having the image url for the vehicle, description column having the detailed description on the type of the car, county column having the county of the car availability, state column having the state of the car availability, lat column having the latitude coordinate of the availability, long column having the longitude coordinate of the availability, posting_date column having the date of posting of the advertisement.

1. Data Pre-processing and EDA

For this phase of the project, we are dropping few unnecessary columns and considering only the columns which we feel will be suitable for the price prediction and the time series analysis. So initially we did an imputation of the year, odometer, lat, long columns to fill in the missing values as the number of missing values in these columns were less and it seemed reasonable to impute. We used KNN Imputer. Also some of the columns had null values for more than half the number of records, so we had to drop them (rows like id, url, region_url, size, county, image_url, condition, cylinders, VIN, drive, paint_color). So, the final cleaned data set had around 14 columns, mostly focusing on the features of the vehicle. Also, we cleaned up the outliers from the price and odometer columns which will later impact the accuracy of the model. Price column had extreme outliers, which were definitely noise and had to be removed.

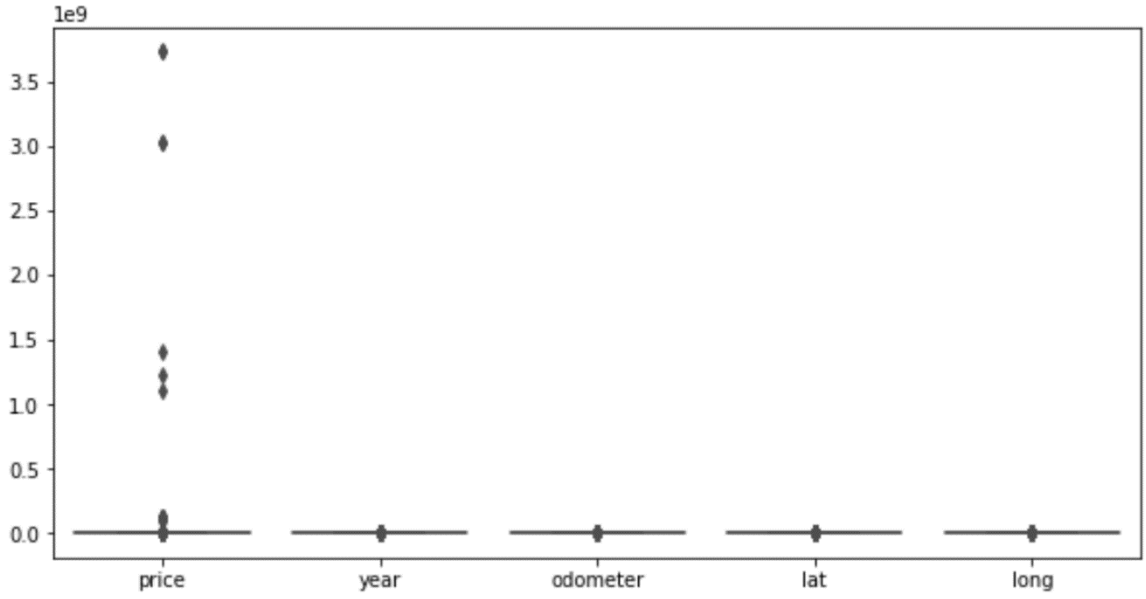


Fig 1: Box plot to check outliers

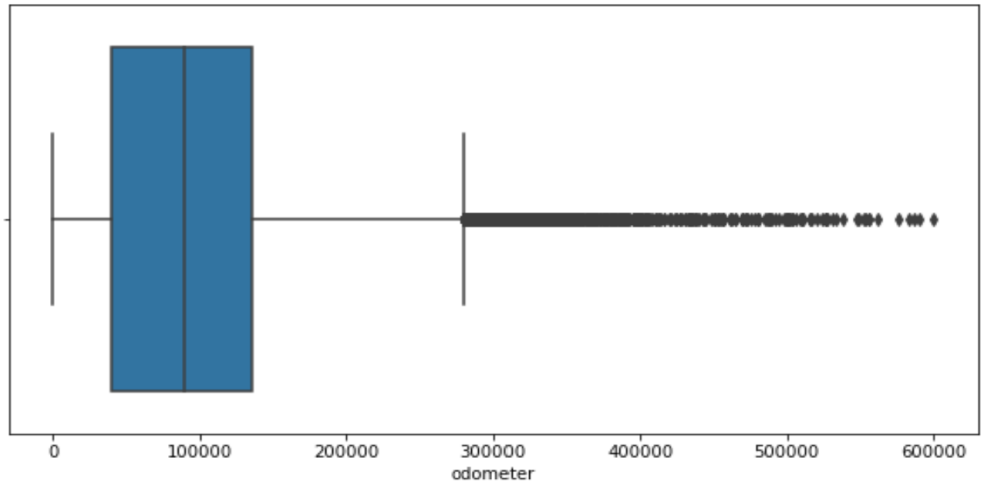


Fig 2: Box plot showing the number of outliers in the odometer data

After data pre-processing, we performed some exploratory data analysis to see each feature's impact on the price. We wanted to see how the price has been varying over the years for which we plotted a line chart as below. This graph was plotted for the whole of the data. As seen from the graph, there are some outliers (year 1900), which were cleaned before modeling.

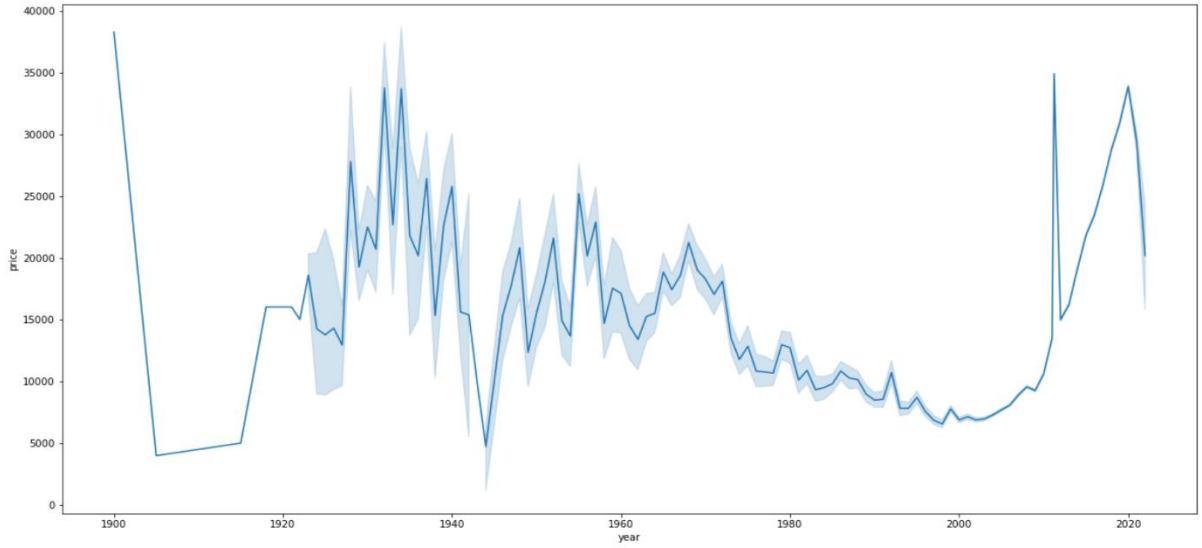


Fig 3: Variation of price over time

To see how the price value is distributed based on car manufacturer, we plotted a barplot as below:

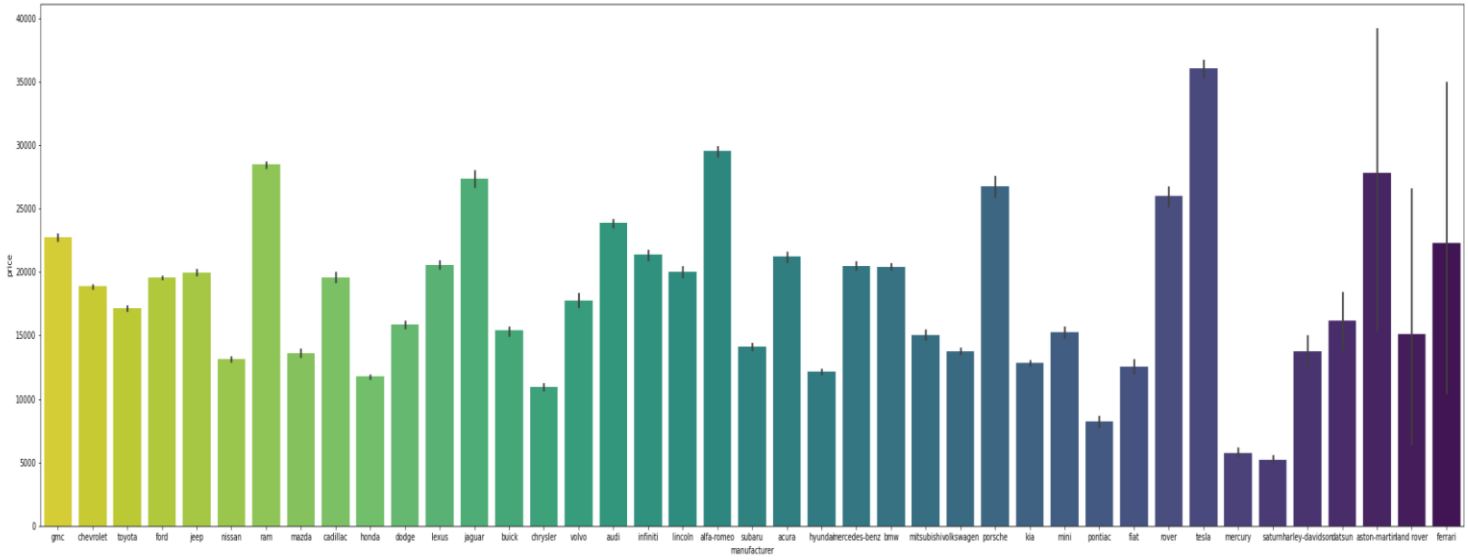


Fig 4: Impact of car brand on price

To see how the price values are impacted by the fuel type we plotted barplot as below:

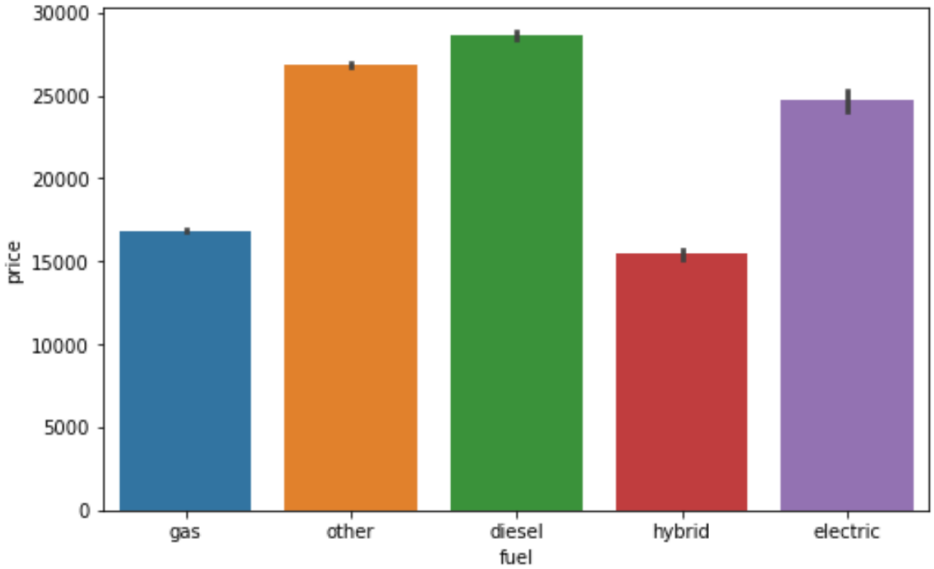


Fig 5: Fuel vs Price

To see how the odometer reading impacts pricing of the car, we plotted a scatter plot which will give us an insight into how the price reduces with high odometer reading.

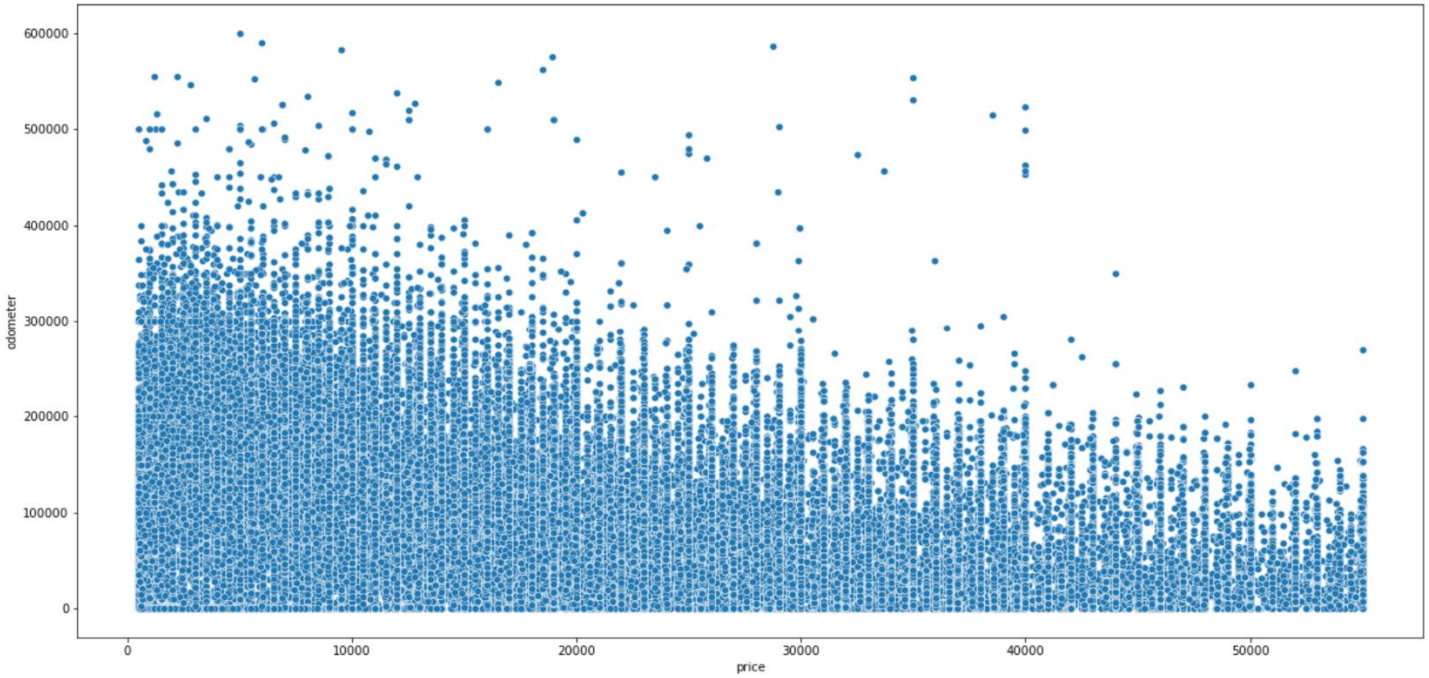


Fig 6: Odometer vs Price

To gain more insight from the data, we plotted bar chart showing how the price is impacted with the transmission type, where we could see other(hybrid) costs the most.

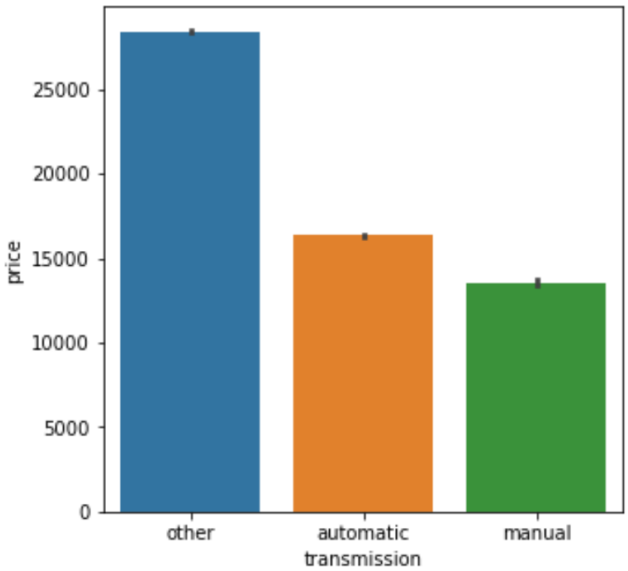


Fig 7: Transmission vs Price

Also, in order to see the variation in price based on car type, we plotted a barplot with type on the x-axis and the prices of the types on y-axis.

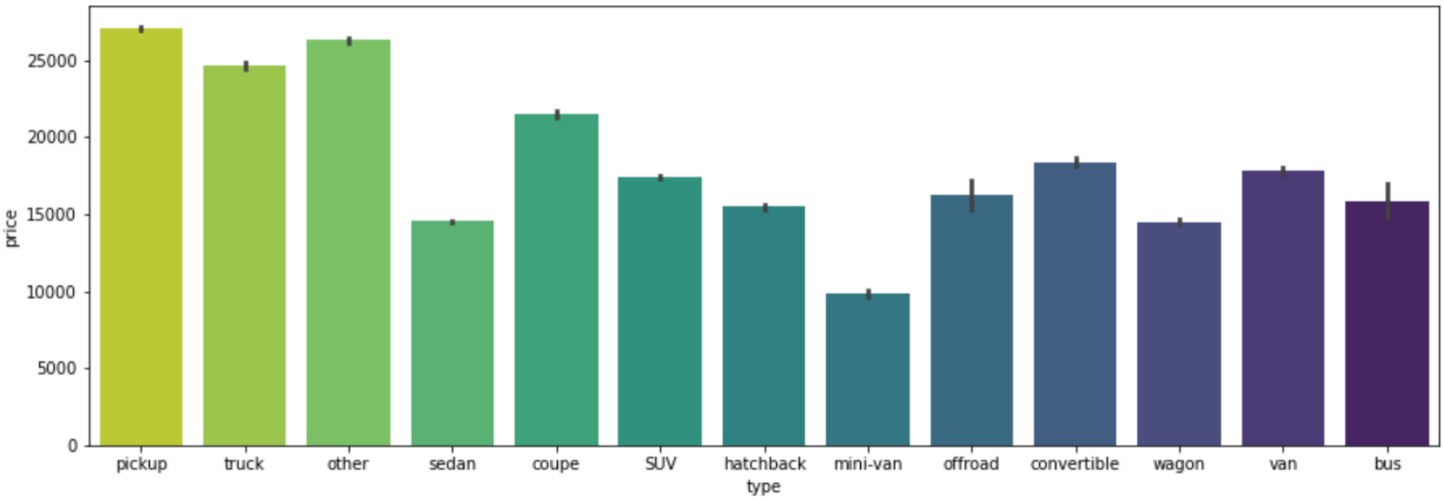


Fig 8: Car type vs Price

To see the variation in price with the status of car (whether it is clean or having any missing part) which can impact the price by a huge amount we plotted a barchart as below.

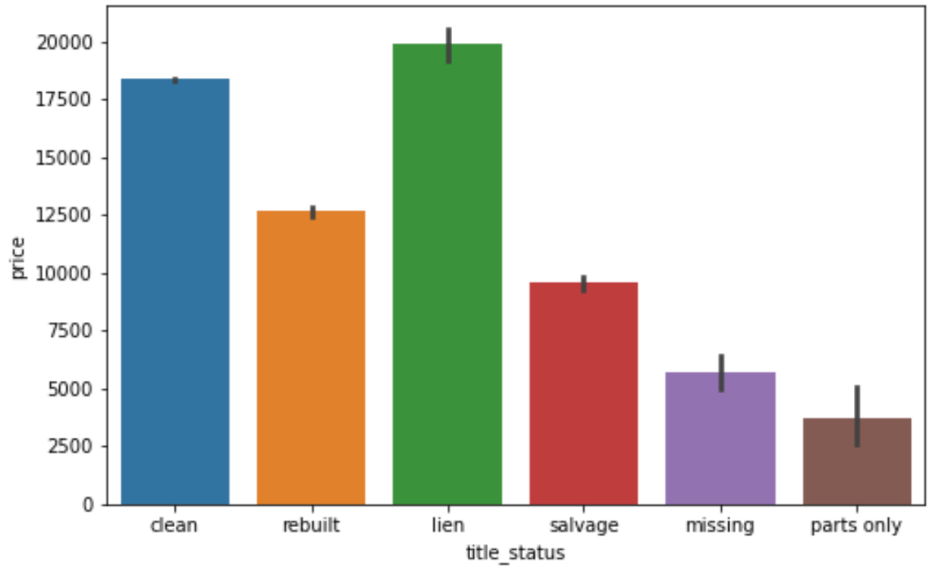


Fig 9: Car condition vs Price

To see how the price of the used vehicle varies across various states of US, we plotted another barplot with states on x-axis and price on y-axis as below:

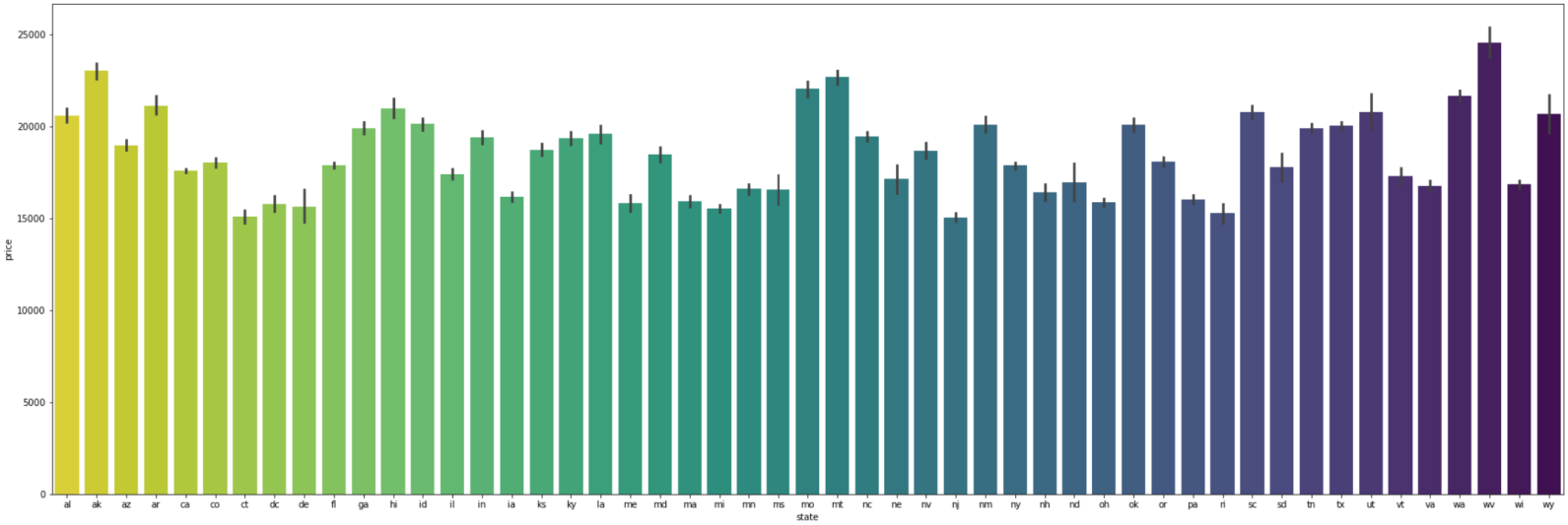


Fig 10: States vs Price

Now finally in order to understand the correlation of the price column with the other columns, we plotted the following correlation heatmap:

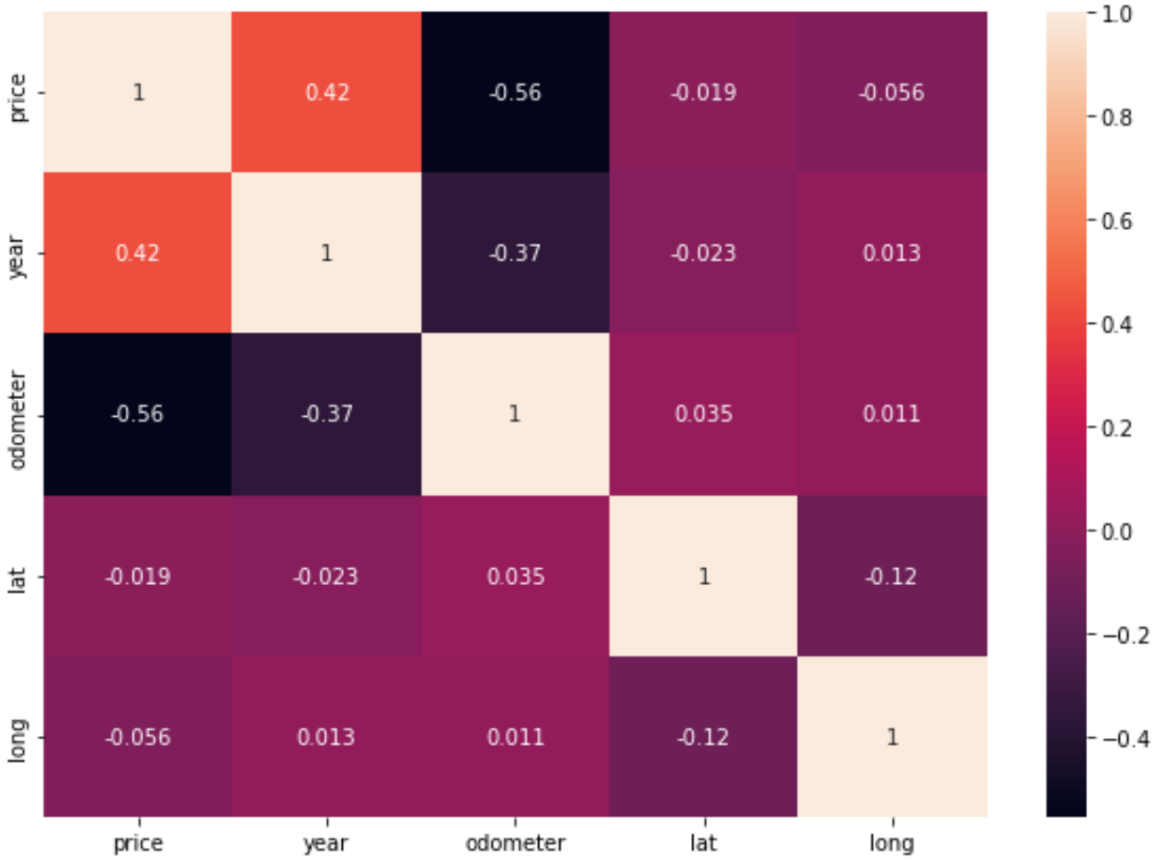


Fig 11: Correlation Heat map

2. Feature Engineering

During the final phase of the project for improving the efficiency of our model we performed certain feature engineering. We created a new column for the age of the vehicle as the build year of the vehicle is highly correlated with price. For this calculation we took the current year and subtracted the build of the car which gave us the age of the car. The next column which we created is miles travelled per year, as miles covered is another field which is highly correlated with the price. We dropped few of the columns which does not significantly affect the price of the vehicle. After adding these new columns, we performed label encoding to some of the categorical values. After the label encoding we converted the categorical features to indicator variables using the get_dummies function is python. Once the feature engineering is performed we took the highly correlated features which can be used to train our model. The below are the features which we planned on considering:

```
price                      1.000000
odometer                   0.555178
age_of_car                 0.424100
made_year                  0.424100
transmission_other         0.365170
transmission_automatic     0.277910
fuel_gas                   0.267389
type_sedan                 0.249397
type_pickup                0.247329
fuel_diesel                0.198051
fuel_other                 0.192959
type_other                 0.151555
type_truck                 0.146147
mile_travelled_per_year    0.094096
title_status_clean         0.090711
transmission_manual        0.089483
type_mini-van              0.074605
title_status_salvage       0.069383
title_status_rebuilt       0.061285
type_coupe                 0.058615
type_wagon                 0.049157
manufacturer               0.048865
type_hatchback             0.044491
title_status_missing       0.036716
fuel_electric              0.033581
type_SUV                   0.028838
fuel_hybrid                0.025080
region                     0.021982
title_status_parts only    0.020037
title_status_lien          0.008194
type_offroad               0.006019
type_bus                   0.005155
type_van                   0.004154
type_convertible           0.002280
Name: price, dtype: float64
```

Data Analysis

We have used machine learning models to enable future prediction of the selling price of used cars. We have used regression models since the price column, which is to be predicted, is continuous and numerical. Region, year, manufacturer, fuel, odometer, title_status, transmission, type, state are the input columns used. The decision to keep these specific columns were made after checking the feature correlation with the output and also after taking into account the missing values.

1. Modeling

Models used:

a) **Linear Regression**
b) **Random Forest**
c) **XGBoost**
d) **LightGBM**
e) **OLS**
f) **Deep Neural Network**

We decided to use Linear Regression as it would best serve as a baseline model and we aimed to further improve the prediction by using other better models. We then used Random Forest as it is more advanced and is known yield better results. The other two models used were XGBoost and LightGBM, as these models work effectively on a large data set and are robust to outliers. Another main reason for choosing these models was that all these models allow n_jobs or parallel processing. Since we have close to 360000 rows of data, without parallel processing it would be very time consuming. OLS model was used separately to understand the basic standard error in the model.

The results obtained using the above models:

Model,R2 Score		RMSE
Linear Regression	0.409897	9556.57
Random Forest	0.853838	4756.15
XGBoost	0.77662	5879.77
LightGBM	0.746814	6259.76

Fig 12: Model Results

Random Forest is observed to be the best model and all these results are obtained without any hyper parameter tuning.

There is no data sub setting or sub group analysis. We decided to use the whole data after cleaning. We will be implementing cross fold validation in the next phase. There were two statistical tests run as follows

Tensorflow based DNN Model:

We also implemented a Deep Neural Network model for the price prediction using tensorflow. The model has 2 neural layers and we used ReLU as the activation function. The below is the summary of the DNN model:

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 normalization_7 (Normalizat  (None, 9)                3
 ion)

 dense_33 (Dense)            (None, 64)                640

 dense_34 (Dense)            (None, 64)                4160

 dense_35 (Dense)            (None, 1)                 65

=================================================================
Total params: 4,868
Trainable params: 4,865
Non-trainable params: 3
_________________________________________________________________
```

Fig 13: DNN Model layers

We trained the model with a validation_split = 0.2 and epochs = 100, we were able to get a RMSE of 7179.4316 and loss of 4881.1464 using this model. So the Random Forest model was performing better than this DNN model, so opted to go ahead with further fine tuning of the data using Random Forest model.

Fine Tuned model:

In order make our model more accurate we had to cut shot our data. We decided on checking how COVID-19 has impacted the used vehicle sales, for which we separated our data into 2 section pre_covid and post_covid, where we took all the cars with made year after 2017 to 2019 as pre_covid data and vehicles with made after 2019 as post_covid. From our analysis we were able to conclude that the average car resale price for two years before and after COVID, it can be clearly

seen that it is significantly high after COVID, this was just for an observation how COVID has impacted the used car sales market.

Since we need to predict the prices of the current market, we reduced our dataset on basis of year column. We filtered out the year of made between 2010 and 2020, which can give more idea on the market trend.

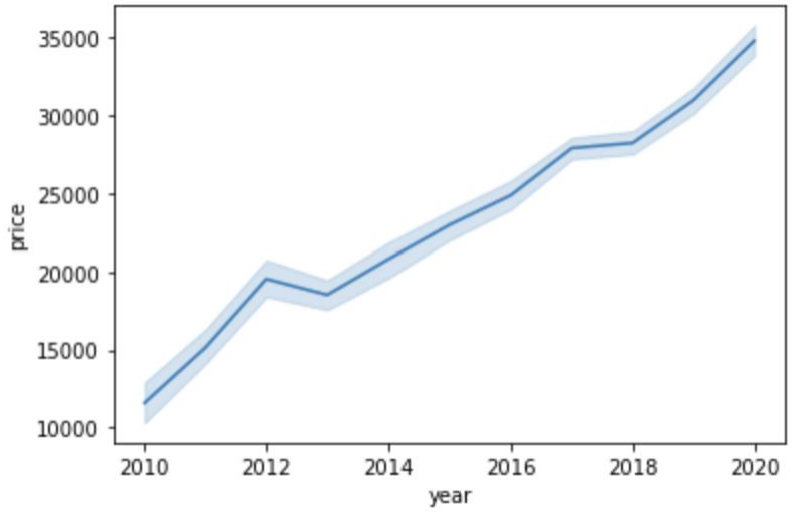


Fig 14: Used Car Prices from 2010 to 2020

Trained our model using this filtered data and were able to get better results compared to the complete dataset. We got a mean absolute percentage error score of 0.21098 which is way better than the previous scores. Also the RMSE was reduced to 5844.812 and R2 score was 0.70179

2. Hypothesis testing:

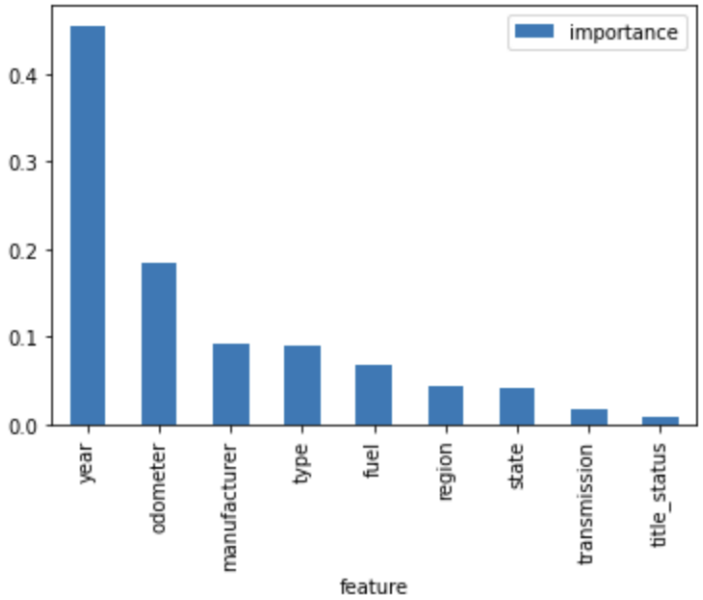


Fig 15: Feature Importance

When feature importance was checked using Random forest, which was our best model, we observed that ‘year’ feature was the most important. So, we decided to perform a hypothesis testing

We performed the testing to understand if older the car is, lower will be the selling price. So

H0: Older the car was, lower the selling price

H1: Selling price is not affected by how old the car is

We kept the level of significance at 0.05 or at 95% confidence interval.

We used Pearson correlation coefficient for this testing

The results obtained were:

Pearson correlation coefficient = 0.424

p-value = 0

From the results, the null hypothesis had to be rejected and also the correlation between the year and price column was moderately high.

3. Uncertainty testing:

Used scipy based OLS regression model to understand uncertainty through standard error.

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.180
Model:	OLS	**Adj. R-squared:**	0.180
Method:	Least Squares	**F-statistic:**	7.813e+04
Date:	Thu, 31 Mar 2022	**Prob (F-statistic):**	0.00
Time:	16:44:26	**Log-Likelihood:**	-3.8283e+06
No. Observations:	356245	**AIC:**	7.657e+06
Df Residuals:	356243	**BIC:**	7.657e+06
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.216e+06	4413.836	-275.392	0.000	-1.22e+06	-1.21e+06
x1	613.3708	2.194	279.510	0.000	609.070	617.672

Omnibus:	55789.412	**Durbin-Watson:**	1.533
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	93957.475
Skew:	1.046	**Prob(JB):**	0.00
Kurtosis:	4.398	**Cond. No.**	4.71e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.71e+05. This might indicate that there are strong multicollinearity or other numerical problems.

It can be seen that the standard error for the data is around 2.194 and is improved by using advanced models.

4. Hyperparameter tuning:

From our analysis on the last phase, the Random Forest model was performing the best on our dataset. We decided to perform hyper parameter tuning on this model to find the best parameters for our model. We used the metrics like R2 score, RMSE (Root Mean Square Error) and MAPE (Mean Absolute Percentage Error) to choose the best model. For the hyper parameter we used the randomized search and the parameters like max_depth and n_estimators with range of values. The best parameters which we got for our model are:

```
{'n_estimators': 600, 'max_depth': 80}
```

We trained our model using the best parameters and below is the metrics score which we got for the model:

Metrics	Random Forest Model
MAPE	0.4309
RMSE	4734.55
R2	0.8552

To fine tune the model we used more parameters with randomized search and below are best parameters which we got during this process:

```
{'n_estimators': 600,
 'min_samples_split': 2,
 'min_samples_leaf': 2,
 'max_features': 'log2',
 'max_depth': 80,
 'bootstrap': False}
```

We again trained our model using the best parameters and below is the metrics score which we got for the model:

Metrics	Random Forest Model
MAPE	0.4475
RMSE	4792.40
R2	0.8516

Argument:

Key Arguments from data processing/analysis:

1. Random Forest is the best model to predict the used car price. This is evident from the r2 score and rmse value comparison in the data analysis section. This model is able to explain close to 86% of the variance.
2. ‘year’ is the most important feature and from the hypothesis test, it can be told that older the car does not mean that lower will be the price.
3. Odometer has the highest correlation with the price to be predicted. This can be seen from the correlation plot.
4. Selling price is high in states like Alaska, West Virginia, Montana and low in states like Rhode Island, Connecticut and New Jersey. This can be seen from the ‘price’ vs ‘state’ plot.
5. Pickup trucks and diesel vehicles have the highest selling price as seen from the plots.
6. From the time series plot of how the price changed over time, it can be seen that selling price used cars dropped till early 2000 and rapidly increased around the year 2003 and is still increasing.
7. From the odometer plot, it can be seen that more distance a car is run, lower is the selling price.
8. Comparing the average vehicle resale price for two years before and after COVID, it can be clearly seen that it is significantly high after COVID
9. There has a been significant increase in the price of the used cars from 2010 to 2021 and the trend is forecasted to continue.

Design

Intervention

Our final aim with this project was to design a model which can predict the used car prices using any dataset. We also wanted to create visual dashboard, which is enabled by our robust machine learning model, which would show our stakeholders the price on the basis of the features which they selected. The target behavior of our intervention was to predict accurate prices of the vehicles based on the selections of features and we were able to achieve this to an extent. The expected behavior of our intervention was to have an option for our stakeholders to select the type, based on which the results and price will be predicted. Our form of delivery for this intervention will be a dashboard which will have the option to choose the type of stakeholder in a drop down and then add the features required and based on these parameters the price will be predicted and displayed as such.

To measure the effects of our intervention we have the following metrics and based on these metrics the necessary changes will be performed:

1) Accuracy of the model
2) Perform a stakeholder's survey or feedback

Design

When it comes to used car sales, it can be seen as the derivative of the new cars in the market, and this is the same all across the globe. When we are taking the case of a country like United States, around 80% of the newly sold cars are either leased or financed and the same will be available in the market once this period is over. So, we decided to build a machine learning based platform which will help predict the price of the vehicles based on various features and can be used by all the stakeholders.

There are many stakeholder group that would be interested in a used car price prediction model. But we are mostly focusing towards three groups of people with our model. Customers/Buyers are the people who are looking to get a car for their usage. Usually if a customer needs to get buy a car, the first thing which they will be doing is that they will search the same on the internet which

will in fact give them a detailed idea on the pricing for the car model which they are looking for, but it is still very difficult without proper research on this. And if the customer is a person who does not have much time to spend on this research, he/she might go and approach an agent for buying a car, but later they will have to pay extra for this service which usually depends on the model and conditions of the car. But what if you can know the buying price of the car even without the help of an agent and directly get in touch with the seller.

The second group of stakeholders which we are targeting are the sellers, the people who may wish to sell his/her car to upgrade from an old model to a new one or people who are in need of some urgent money for any other purpose. Usually, these people will want to sell their car as soon as possible for which they will not be able to research much on the price. There are also chances that they approach an agent and end up paying some amount for this service. This is where our model comes into picture. The model will help these people to get an estimate price for the car they wish to sell based on the features, manufacturer, model etc. The main motivation of a seller is to sell his car with a reasonable rate according to the market price. But without a proper understanding of the market price and worth of the car, the person might end up in selling his car for even lesser price.

The third group of stakeholders which we are targeting are the small-scale dealers who would like to start a new venture of selling used cars. A model like this will really help to analyze the current market trend and the time series data will help them analyze how the price varied across time. These group of people usually follow a method like they will buy the cars from the buyers and sell this by acting as an agent in between the buyer and seller, so a detailed idea on the price may help them to setup a good deal. The main motivation for these group of people is to extend their business and for this they need the deals taking place.

After the midterm, we conducted a walk-through usability test with potential stakeholders and iterated our designs based on their feedback. We presented our idea and talked to 3 various stakeholders, and below are the insights we gained.

1. Subject knowledge:

The stakeholders when they try to use the model, they do not need any prior knowledge and idea about the model or the dashboard. The interface should be fairly simple to allow easy understanding by anyone

2. Easy to access:

Stakeholders found the model to be easy to access and they just had to choose the parameters on the dashboard which will give them the predicted price. Any person who is given the dashboard should be able to access it with ease. The feature options should be available in a very accessible way.

3.Meeting the expectation:

Every stakeholder, when hearing of such a platform and its scope, will have some basic expectations. Those expectations were noted and diligently incorporated into the platform design, aiming to better the user interaction and retention. Any of the future feedbacks will also be incorporated.

The information gained from the survey is highly useful for the design. The information generated from the platform can be shared through a digital or physical space, where the insights gathered can be saved in a excel file or printed as a report. And there is not much concern about the sharing or transferring of this data as it can be done through any communication systems, either through an email as an attachment or through a physical mail. The storage of these documents will either be placed into databases which, on cloud or even on an external hard drive. So, later stakeholders can take this printed or digital report when they want to go ahead with a transaction.

When considering the design requirements that arises due to the context in which our stakeholders will interact with the data, we decided to focus on the efficiency of our model. For any firm not having an in house analysis team, It will require extensive work and research to forecast the price a of a new vehicle model in the upcoming years. So our model can provide the necessary information that they need to analyze which would make their task easier. Also, we would like to focus on the accessibility of the model as it should be clearly made in a format to be read and understood by a anyone without a prior knowledge. This will eliminate the issues with communication within the stakeholders.

Visualization

For our design, we decided to construct a digital dashboard that the stakeholders can use to fetch the necessary target. This design was inspired by the existing business websites of this nature. Below we have presented our digital dashboard with the data which is user friendly and no prior

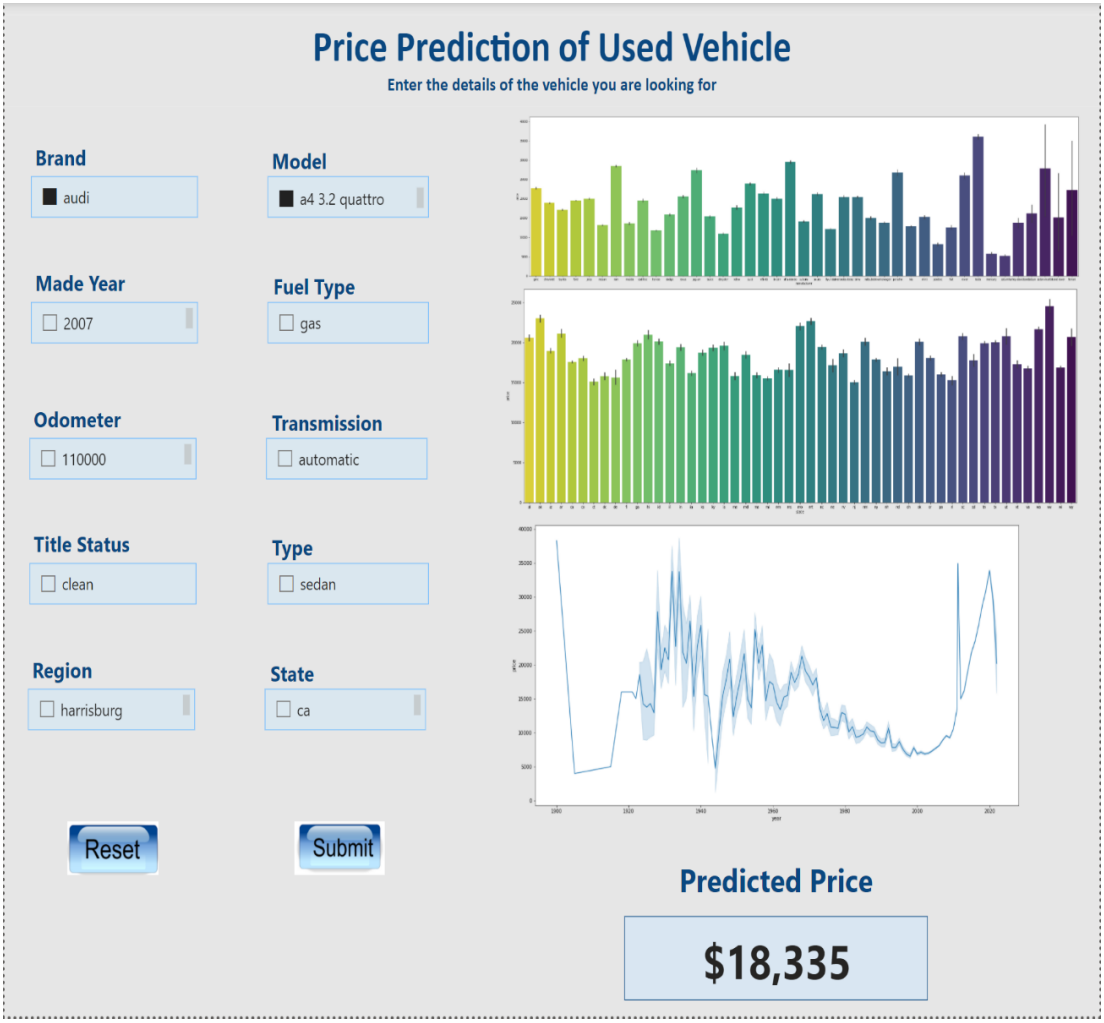


Fig 20: Dashboard Representation

training is required for accessing this dashboard. The effectiveness of the visualization that we are trying to achieve is to show that our model works well (and is getting better) for which we have included the Predicted Price field in the dashboard showing how much is going to be the predicted price. Also, we tried to show the visuals of how price is distributed across the brands to show the top selling manufacturers and the expected price range for them. We also added a bar chart on the dashboard showing how the price is distributed in different states within United States so that a stakeholder can get a rough estimate of the price and these visuals help in comparing with the predicted price. We also have added an effect for the time series analysis of the historical data, like

how the price have been varying over the years, which can help the stakeholder to forecast the price of vehicle in the upcoming years. The dashboard which we have provided is highly user friendly and can be even used a person who is not highly proficient with computers.

As a result of our design, we can easily communicate to our stakeholders the information they need to know in a concise manner, by showing all the information they need in a single place which will make their job easier in the long run. This can also help them identify specific patterns or entities to target. In addition to this, included features facilitate efficient information sharing among audience.

Ethics

1. The values that we focused on were to provide all the three types of stakeholders a fair resource to utilize to predict the selling price of an used car. There will be absolutely no bias over any of the stakeholder. All the predictions made will be purely driven by historic market data.
2. To achieve our value goal, we have compared and checked our data with other car selling platforms and current market data. We ensured that the data we collected is not biased to a particular stakeholder in any way.
3. Some ethical ways in which our platform could be used is, any of the current used car dealers can make use of our platform to fix the price of a car in a fair way, if it was already over priced. There no explicit unethical way in which our platform could be used. In certain cases where the predicted results are inaccurate, the chance of which is very less, there might be some setback for one of the stakeholder.
4. Our design in no way disproportionally affects any of the marginalized group. This equal and fair resource that could be used by any one. One consideration is that, our platform is currently restricted only to the US market and will not be available anywhere else.
5. Our platform is basically about used cars. Details of many cars which are very old will also be present. From the data of manufacture, one can easily figure out the number of old cars by their manufactured year in a given region. This data can be used to scrap old cars, there by driving a positive impact on the environment. Also our platform minimizes the resources taken by any of the stakeholder to find out the price of an used car.

6. Our goal as previously said is to provide a fair and free platform to predict a used car price. With this goal, we could deliver a positive change in the society by eliminating the selling of used cars for high personal profits.

Project code:

Link for the web scraping code: https://github.iu.edu/avgopal/INFO-I-513--Final-Project/blob/main/Data_Extraction2.ipynb

Link for the EDA code: https://github.iu.edu/avgopal/INFO-I-513--Final-Project/blob/main/Final%20Project%20-%20Phase%202.ipynb

Link for the modeling code: https://github.iu.edu/avgopal/INFO-I-513--Final-Project/blob/main/Final_Project_Phase%202_Model%20Selection%20.ipynb

Link for the feature Engineering code: https://github.iu.edu/avgopal/INFO-I-513--Final-Project/blob/main/Phase%203_Feature%20Engineering.ipynb

Link for the final model: https://github.iu.edu/avgopal/INFO-I-513--Final-Project/blob/main/Phase%203_Final%20Model.ipynb

References

1. https://fortune.com/2021/11/01/used-car-prices-high-carmax-2021/
2. https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data
3. https://stackoverflow.com/questions/26050855/getting-uncertainty-values-in-linear-regression-with-python
4. https://towardsdatascience.com/practical-practice-of-hypothesis-testing-on-house-price-dataset-1fb169bc04ee
5. https://towardsdatascience.com/python-statistics-for-beginners-pearson-correlation-coefficient-69c9b1ef17f7
6. http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26612934.pdf
7. https://medium.datadriveninvestor.com/end-to-end-project-on-used-car-price-prediction-3dc412d24aa0