**Capstone Project** **Document**

Data Science and AI Course

Part-time – 2021-05-29

# Uchenna Austin Okoli

# Process overview

Graphical user interface, text, application

Description automatically generated

# Problem statement

* What is the problem or the opportunity that the project is investigating?

Predict the price of used cars.

* Why is this problem valuable to address?

Buyers, car fleet managers, used car dealers and car insurers have to evaluate the ‘right’ price for a used car.

* What is the current state (e.g., unsatisfied customers, lost revenue)?

Insurer needs to know the replacement value of a car.

* What is the desired state?

Identifying the depreciation of used cars because it impacts all subsequent decisions.

* Has this problem been addressed by other research projects? What were the outcomes?

A lot of predictions have been done by some individuals in the past. For

example:

Graphical user interface, text, application

Description automatically generated

# Industry/ domain

* What is the industry/ domain?

The Automotive industry.

* What is the current state of this industry? (e.g. challenges from start-ups)

Used car sales are expected to increase 2% in 2021.

* What is the overall industry value-chain?

Extended test drive > Remarketing infrastructure

* What are the key concepts in the industry?

With many consumers [choosing used cars over new cars](https://www.coxautoinc.com/news/gently-used-posing-harsh-threat-to-new-car-sales/) because of affordability, dealerships have an opportunity to focus on their used car programs to help boost their profits.

* Is the project relevant to other industries?

Not really.

# Stakeholders

* Who are the stakeholders? (be as specific as possible)

Insurance companies to know the value of a used car, dealerships, and private sellers

* Why do they care about this problem?

Buyers, car fleet managers, used car dealers and car insurers have to evaluate the ‘right’ price for a used car.

* What are the stakeholders’ expectations?

Some stakeholders in the healthcare sector have conflicting interests. However, the most important stakeholder is the patient, which expects the get the right care at the right time.

# Business question

* What is the main business question that needs to be answered?

Predict the price of used cars.

Can we estimate selling and buying prices for used cars in an affordable and quick way using predictive models based on available data?

* What is the business value of answering this question? (quantify value and make necessary assumptions)

There is a need for a used car price prediction system to effectively determine the worthiness of the car using a variety of features.

* What is the required accuracy? What are the implications of false positives or false negatives?

There isn’t as it is not a classification problem.

# Data question

* What is the data question that needs to be answered?

What is the actual market value of both buying and selling of used cars?

* What is the data required to answer the question?

Data with relevant features that would affect the outcome of the predictions.

# Data

* Where was the data sourced?

Kaggle.com

* What is the volume and attributes of the data?

427 Rows 15 features.

* How reliable is the data?

Reliable.

* What is the quality of the raw data?

Very good.

* Is this data available on an ongoing basis?

Yes. But it is a private data set.

# Data science process

## Data analysis

* What data pipeline was to wrangle the raw data?

Remove unnecessary features and investigated number of null points and percentage of null data points for that feature.

* What are the highlights of the Exploratory Data Analysis (EDA)?

There was a huge difference between both the maximum value/minimum value and the percentiles for some of the features. Which was an indicator of the presence of outliers, which can greatly hinder the performance of our model.

Chart, scatter chart

Description automatically generated

* Is the pipeline reusable? (for example, to process future data?)

Yes.

## Modelling

* What are the main features used?

Model, Make, Type, Origin, Drivetrain, Invoice, Engine Size, Cylinders, Horsepower, MPG\_City, MPG\_Highway, Weight, Wheelbase, and Length.

* Is there a subset of features that would get a significant portion of your final performance? Which features?

Drivetrain, Horsepower, MPG.

* How did you select features?

Categorical Variable Encoding.

* What are the models used?

Decision tree regressor.

Linear regressor.

XGBoost.

* How long does it take to train your model?

A few minutes, with this data set.

* What are the tools used? (cloud platform, for example)

Jupyter notebook.

* Which model was selected?

XGBoost

## Outcomes

* What are the main findings and conclusions of the data science process?

Data cleaning process can be more rigorous by making use of more technical information; like filling missing values with indicator that are more meaningful instead of with empty value.

## Implementation

* What are the considerations for implementing the model in production?

Finding out most relevant features that affect Price would be impactful to sales decisions.

# Data answer

* Was the data question answered satisfactorily?

Yes, it was.

* What is the confidence level in the data answer?

XGBoost was the ideal model for predicting the price of used cars with the model that predicted 83,5% of accuracy the price of used cars, given a set of features.

# Business answer

* Was the business question answered satisfactorily?

The prediction outcome can have an enormous value for dealers and individuals when trying to understand how to estimate the value of a vehicle.

# Response to stakeholders

* What are the overall message and recommendations to the stakeholders?

With the analysis of historical data, it is possible to find patterns that lead to accurate results, which can provide comparative advantage before putting a vehicle on sale or buying it on the market.

.

# End-to-end solution

* What is the overall end-to-end solution to use the model developed in the project?

# References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)

Available at: <https://github.com/uchino83/IOD/blob/main/Capstone_Project%20-%20Uchenna%20Austin%20Okoli%20-%20Used_Car_Price_Prediction.ipynb>

<https://github.com/uchino83/IOD/blob/main/capstoneproject_dataset_cars_data.csv>

* What are the resources used in the project? (libraries, algorithms, etc)

Numpy

Pandas

Seaborn

Matplotlib

Plotly

XBoost

WordCloud

Multiple Linear Regression

Random Forest Models

Decision Tree