Predicting The Price Range Of Used Cars

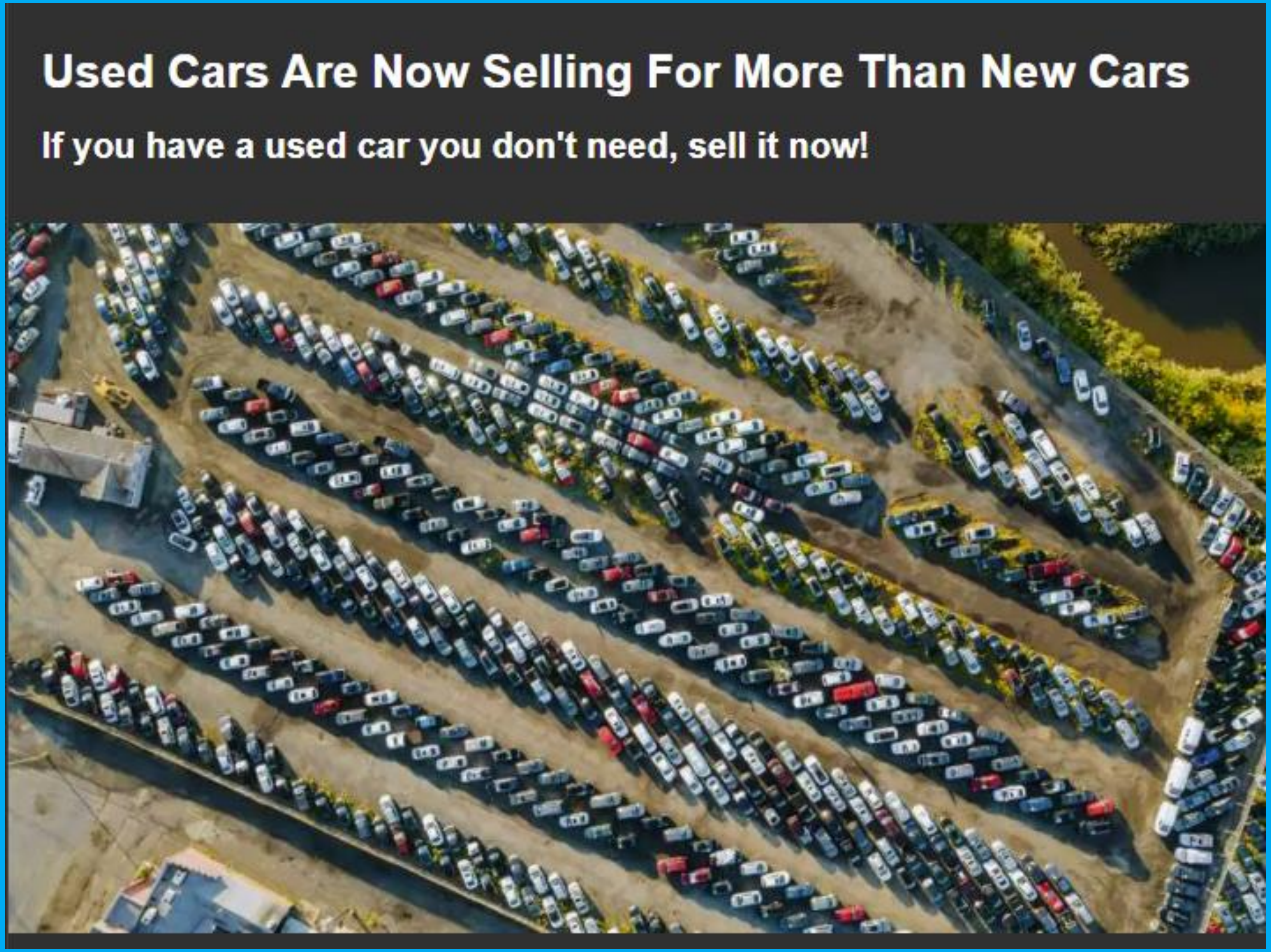
John Ian Castaneda II
501083068

Supervisor:

Sedef Akinli Kocak, Ph.D.

Ryerson University

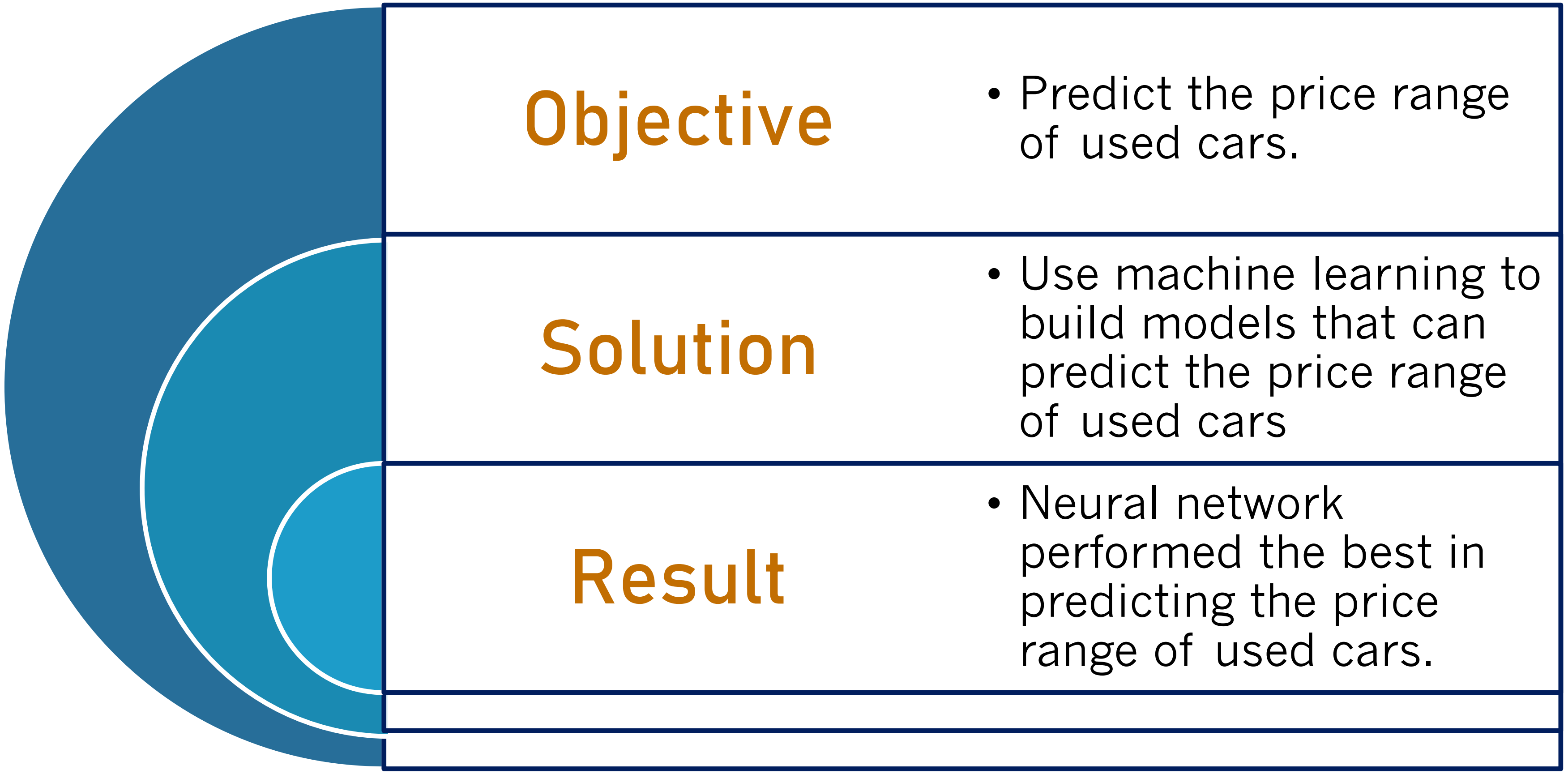
Introduction



Used Cars Are Now Selling For More Than New Cars

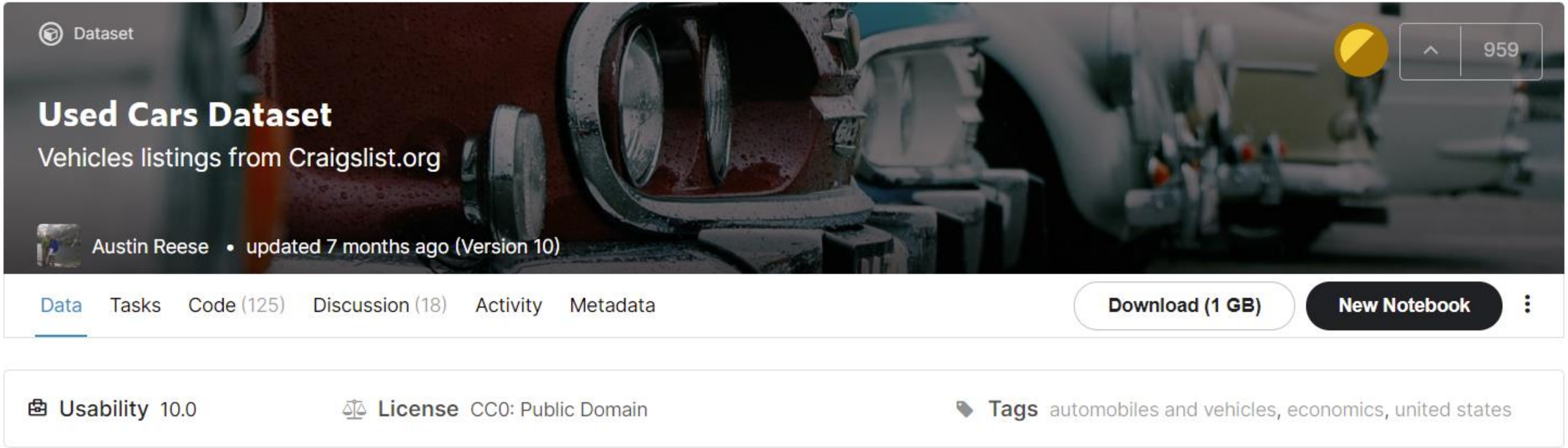
If you have a used car you don't need, sell it now!

Summary



Objective	- Predict the price range of used cars.
Solution	- Use machine learning to build models that can predict the price range of used cars
Result	- Neural network performed the best in predicting the price range of used cars.

Dataset



Used Cars Dataset

Vehicles listings from Craigslist.org

Austin Reese • updated 7 months ago (Version 10)

Usability 10.0 **License** CC0: Public Domain **Tags** automobiles and vehicles, economics, united states

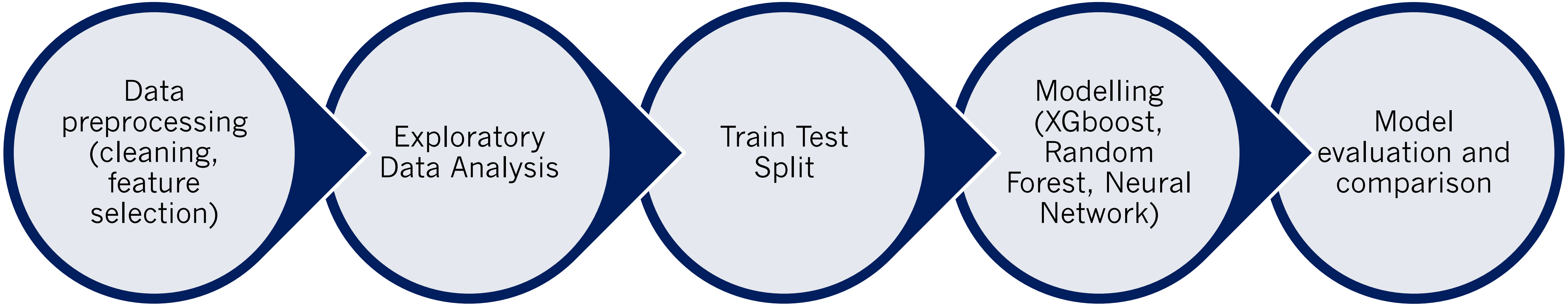
- Dataset is scraped every few months. This was first scraped in 2018. There are 10 versions so far with May 2021 as the latest.
- The dataset has 23 features.

Dataset

No.	Feature	Type
1	url	object
2	region	object
3	region_url	object
4	price	int64
5	year	int32
6	manufacturer	object
7	model	object
8	condition	object
9	cylinders	object
10	fuel	object
11	odometer	float64
12	title_status	object
13	transmission	object
14	VIN	object
15	drive	object
16	size	object
17	type	object
18	paint_color	object
19	image_url	object
20	description	object
21	lat	object
22	long	object
23	posting date	object

- Selected versions are version 2, 3, 8, and 9.
- The combined dataset has 1,961,218 datapoints before cleaning:
  - Pre-pandemic data: 1,076,152
  - Pandemic data: 885,066

Methodology



Data preprocessing (cleaning, feature selection)

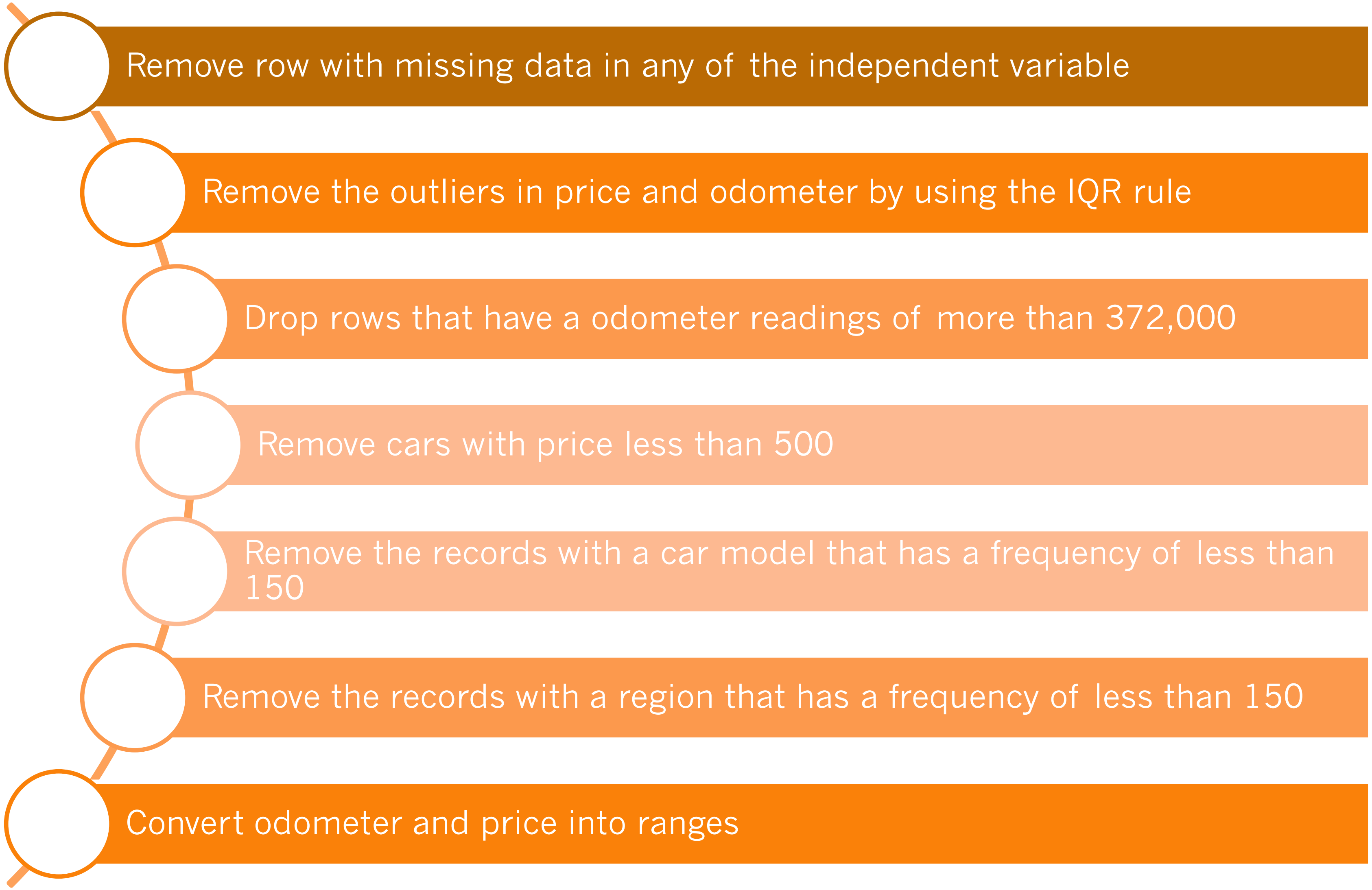
Exploratory Data Analysis

Train Test Split

Modelling (XGboost, Random Forest, Neural Network)

Model evaluation and comparison

Data Preprocessing



Remove row with missing data in any of the independent variable

Remove the outliers in price and odometer by using the IQR rule

Drop rows that have a odometer readings of more than 372,000

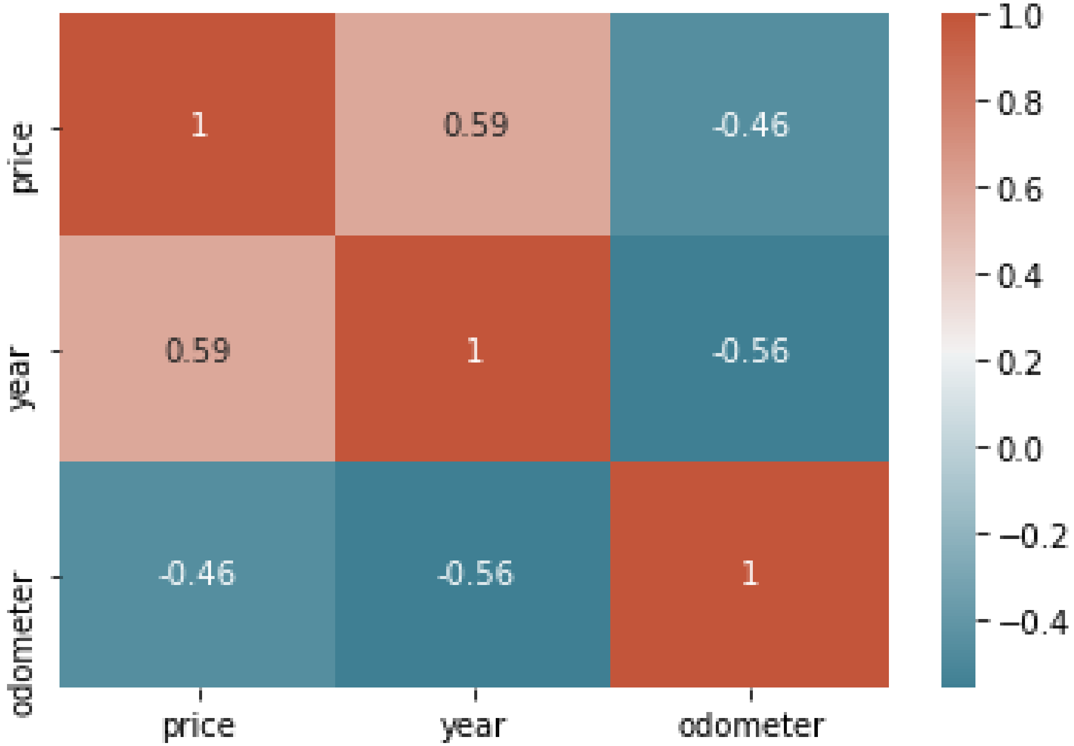
Remove cars with price less than 500

Remove the records with a car model that has a frequency of less than 150

Remove the records with a region that has a frequency of less than 150

Convert odometer and price into ranges

Exploratory Data Analysis

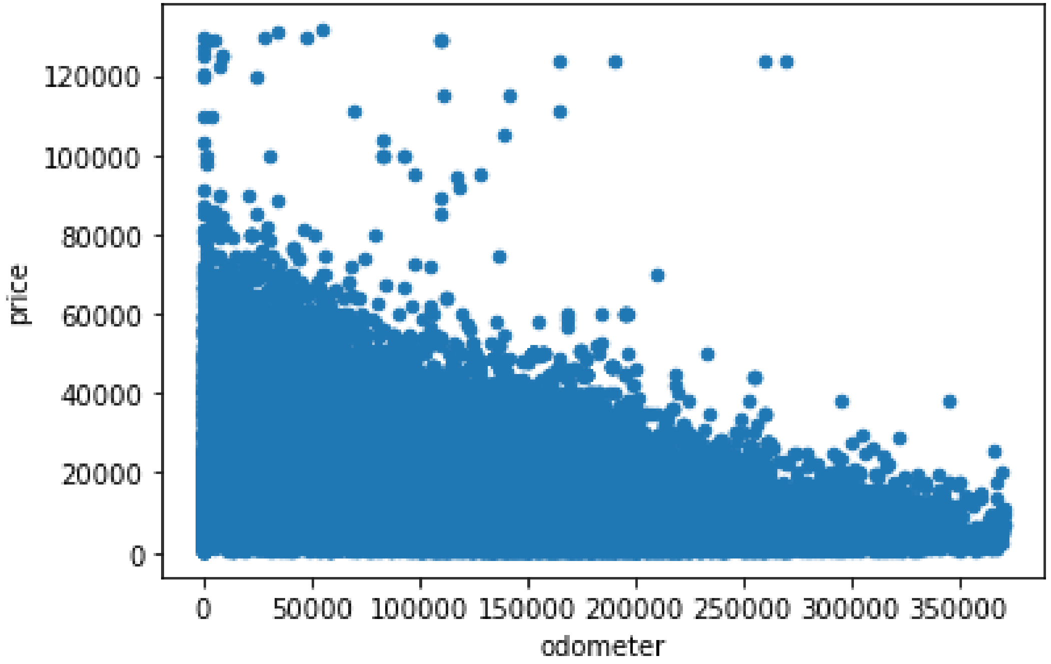


Correlation matrix of price, year and odometer

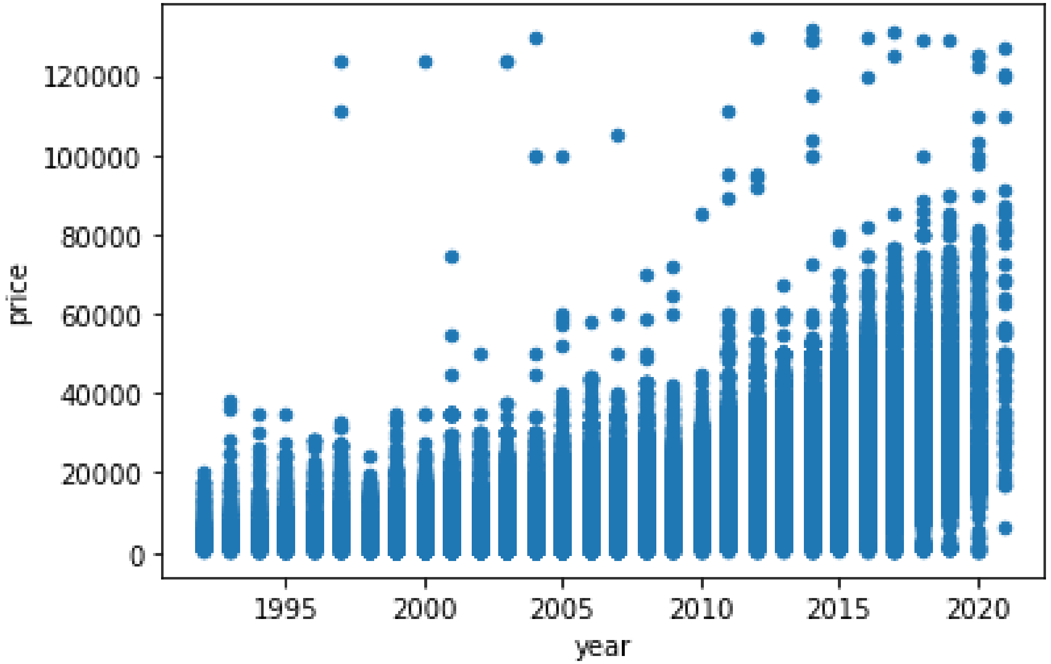
	Price	Year	Odometer
count	194,861	194,861	194,861
mean	12,277.02	2010	118,467.52
std	10,331.43	5	59,912.26
min	505.00	1992	-
25%	4,995.00	2006	77,082.00
50%	8,995.00	2010	116,424.00
75%	16,500.00	2014	156,793.00
max	131,500.00	2021	371,000.00

Descriptive statistics

Exploratory Data Analysis



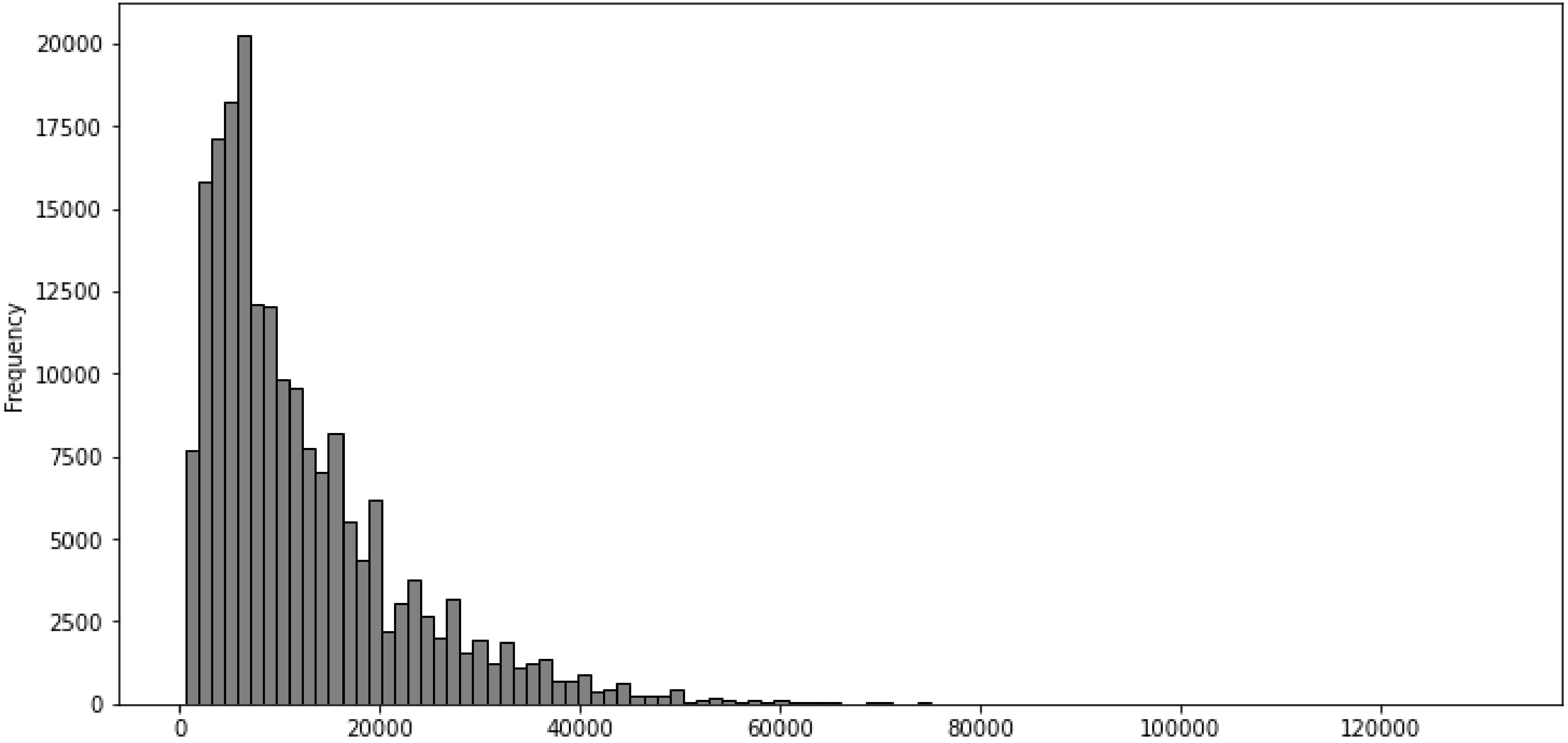
Scatter plot of odometer and price



Scatter plot of year and price

Ryerson University

Exploratory Data Analysis



Price distribution

Results

	Accuracy %	Precision %	Recall %	F1 score %	Cohen's Kappa
Random Forest	76.26	76.13	76.26	76.13	0.6993
XGBoost	67.87	67.60	67.87	67.60	0.5931
Neural Networks	76.67	76.64	76.67	76.60	0.7064

Model performance on pre-pandemic test dataset

	Accuracy %	Precision %	Recall %	F1 score %	Cohen's Kappa
Random Forest	55.31	52.49	55.31	53.58	0.4317
XGBoost	56.20	53.76	56.20	54.72	0.4443
Neural Networks	60.44	58.89	60.44	59.39	0.5032

Model performance on pandemic dataset

Results

	Accuracy %	Precision %	Recall %	F1 score %	Cohen's Kappa
Neural Networks	74.74	74.78	74.74	74.71	0.6834

Model performance of retrained model using the combined pre-pandemic and pandemic dataset.

Conclusion

We were able to predict the price range of used cars using 3 models: random forest, XGBoost and neural network. Neural network performed the best with and F1-score of 76.60% and Cohen’s kappa score of 0.7064.

The models’ performance declined when tested on the pandemic dataset. This is a sign of dataset shift. Retraining the model on the combined pre-pandemic and pandemic dataset resulted to the model having an F1-score of 74.71% and Cohen’s kappa score of 0.6834.

Q&A

Ryerson University