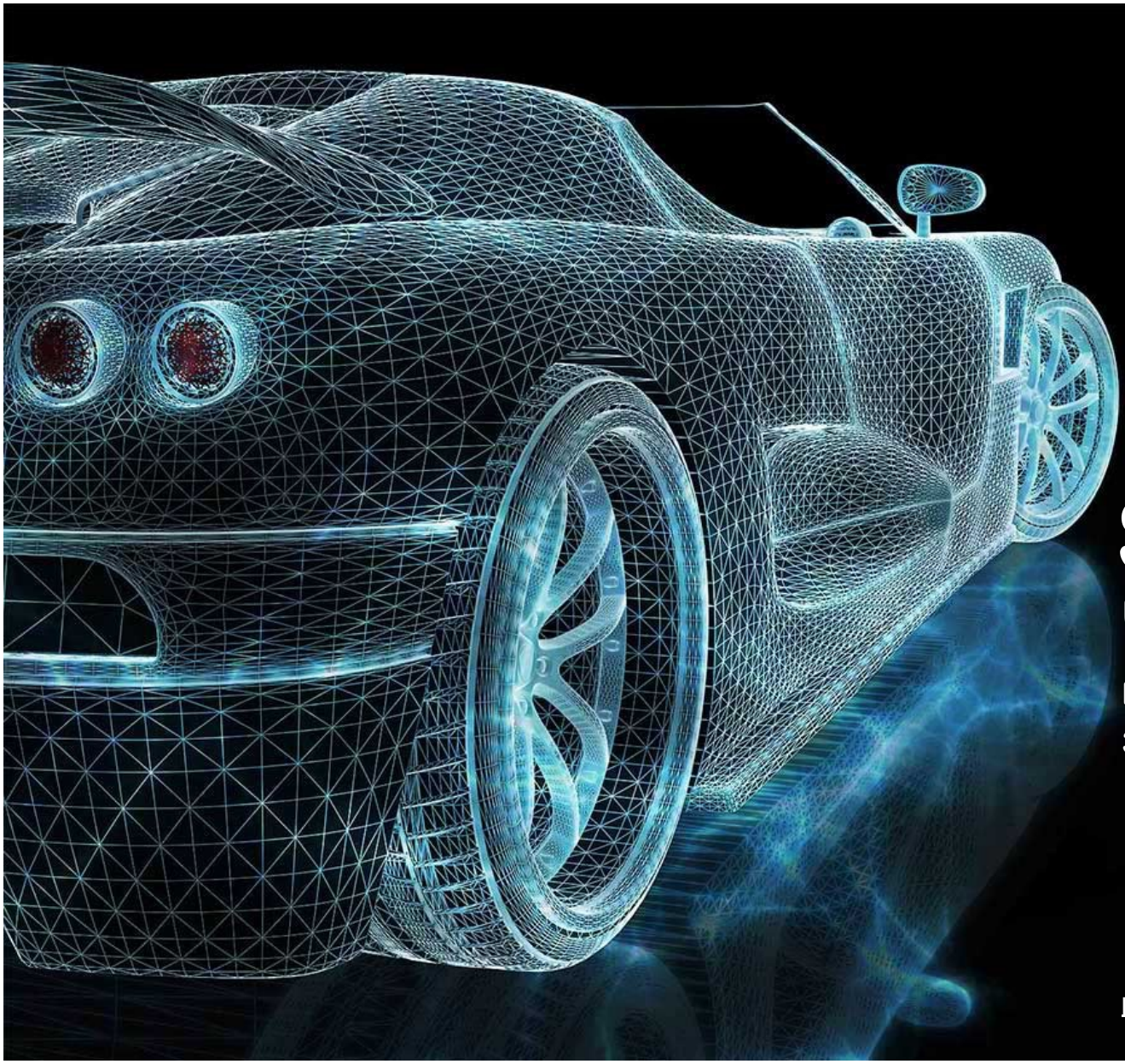


SMARTLIST

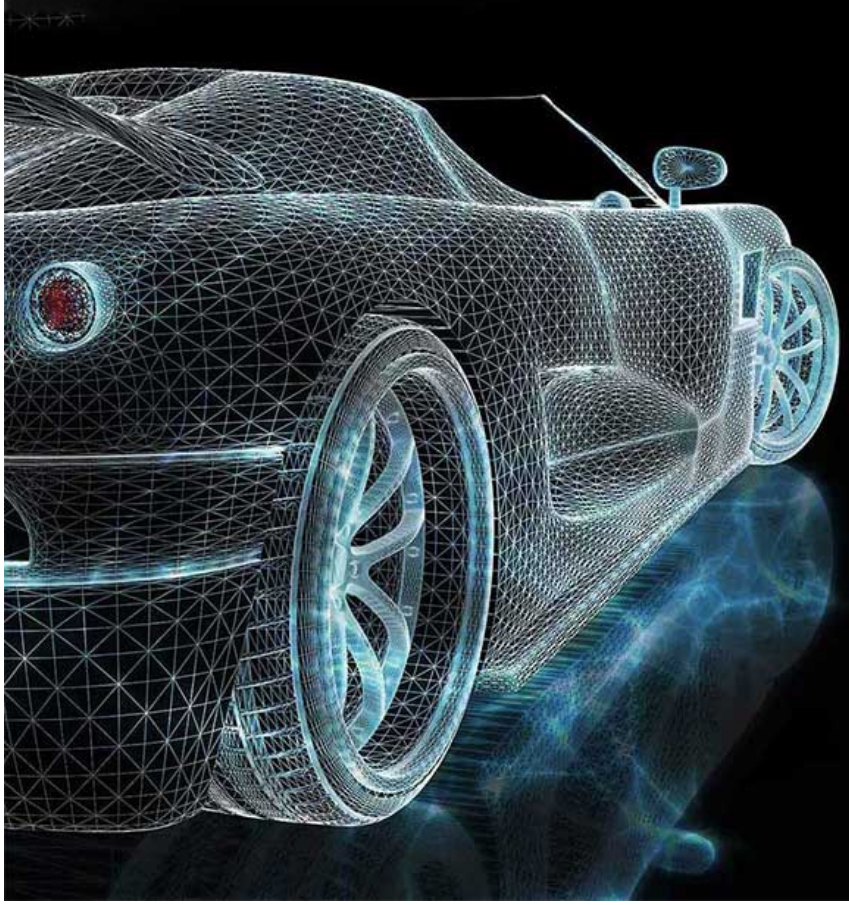
Used Car Price Range Prediction

By Anthony Kwok

30 Nov 2023

[Source Code]

Agenda



SMARTLIST

01 Problem Statement

What Problem | Who Cares | Why We Need

02 Data-Driven Solution

Our Solution | Potential Impact

03 Data & Insights

Dataset | Data Preprocessing | Key Insights

04 Result & Conclusion

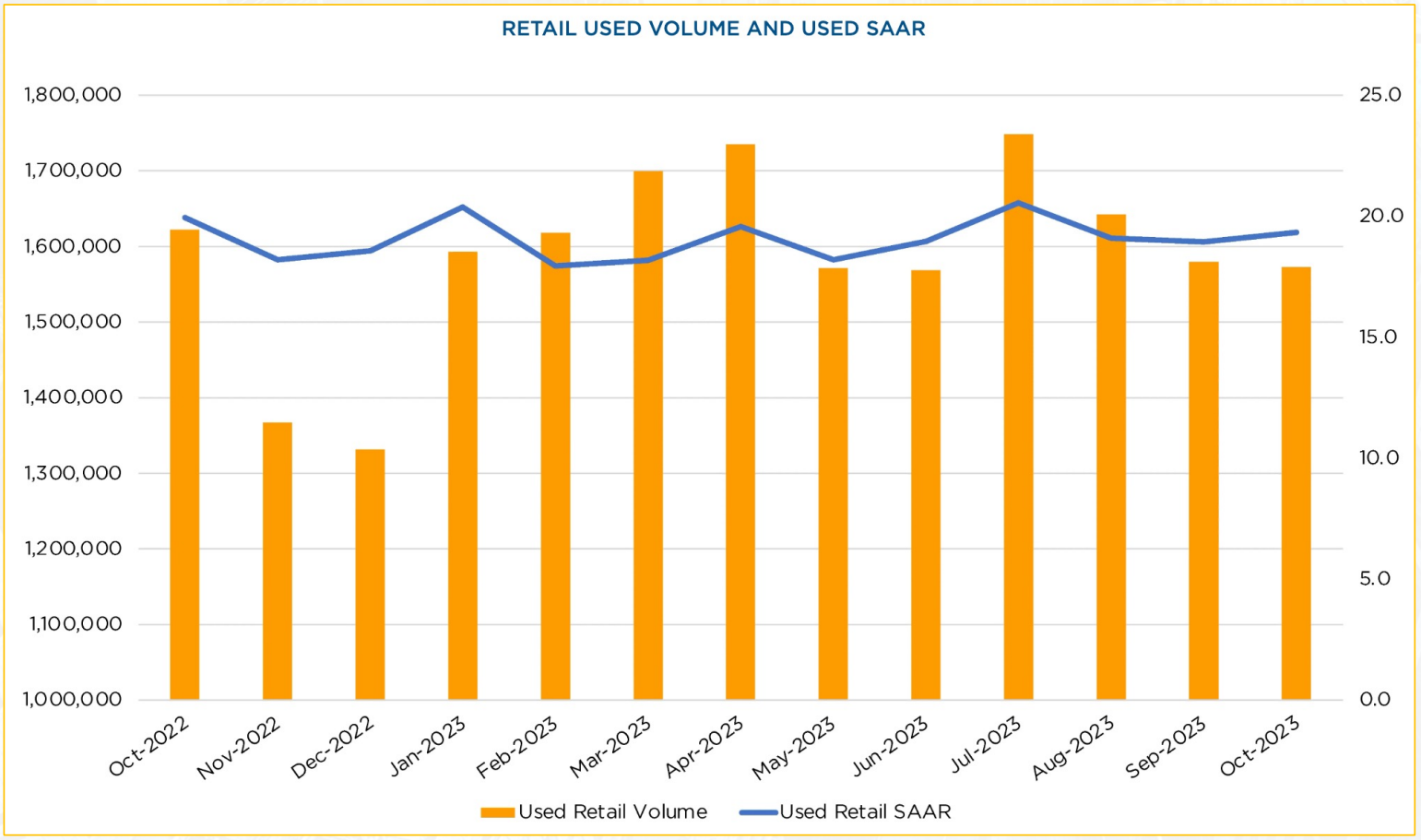
Model Evaluation | Demo | Roadmap

Problem Statement

It takes 4 weeks to sell a used car.

How can we speed up the buying/selling process? To help both buyer and seller to achieve their goal?

Market is Growing!



Monthly Volume

The monthly volume of retail used vehicle in Oct 2023 is around 1.55M ~ 1.75M units.

*Source: https://www.coxautoinc.com/market-insights/estimated-monthly-used-vehicle-saar-and-volume/

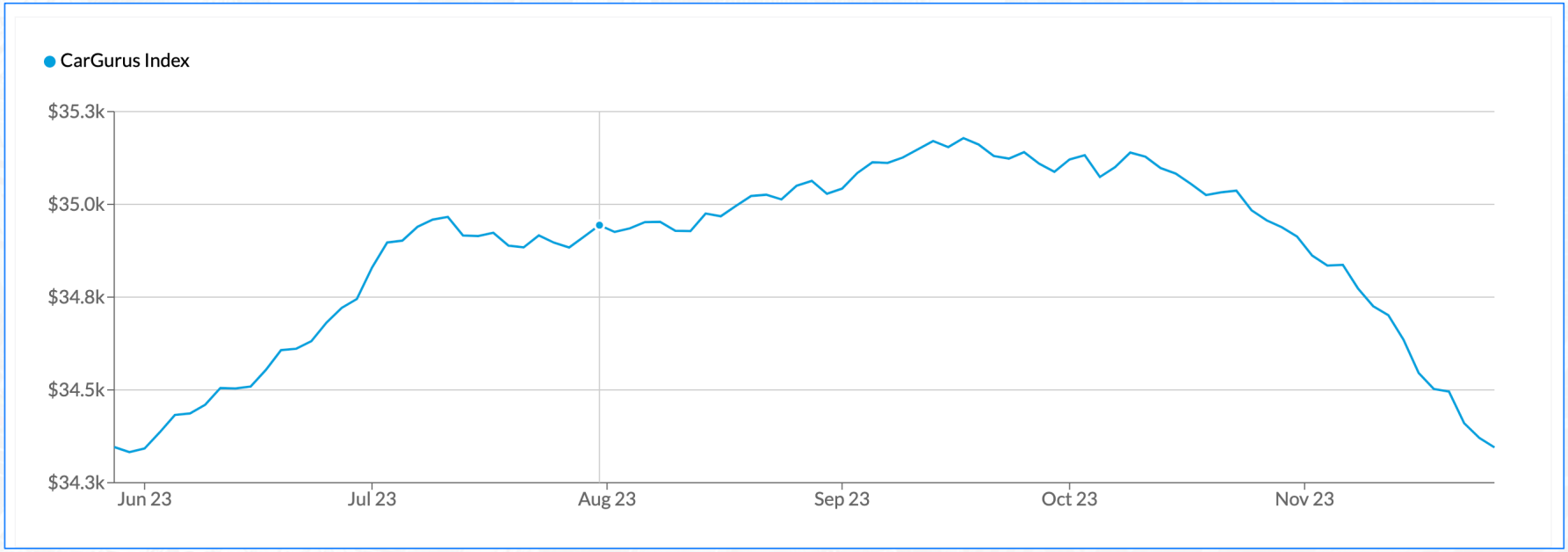
Market is Growing!

SMARTLIST

Average Used Vehicle Price

According to CarGurus Index, The average used vehicle price in Oct 2023 is around USD 35,000.

Total Sales Volume
57.8 Billions



*Source: https://www.cargurus.ca/research/price-trends

Who Cares?

SMARTLIST

Buyer

- Price Uncertainty
- Worry about Overpayment

Dealer

- Hard to manage expectation on both sides
- Long trading time

Seller

- Lack of Knowledge about Vehicle Pricing
- Missed Opportunity

What we need is

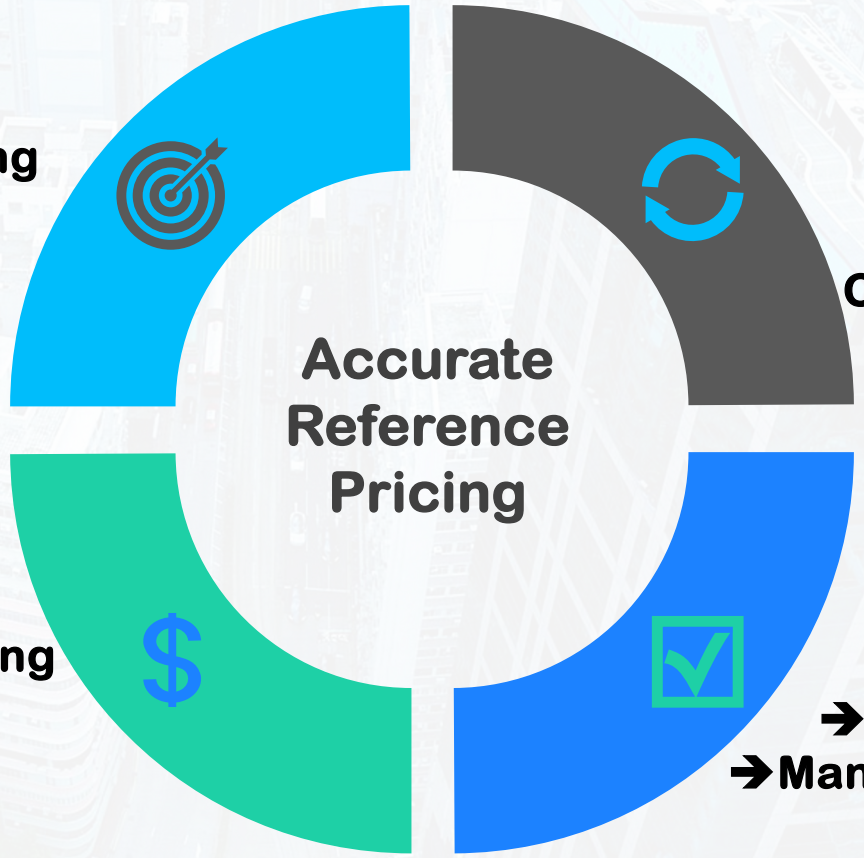
SMARTLIST

Buyer

➔Informed Decision Making
➔Financial Budgeting
➔Accurate Evaluation

Seller

➔Reference Price for Listing
➔Comparative Analysis



Market Efficiency

➔Smoothen the Car Selling Process

Dealer

➔Speed Up Matching
➔Managing Expectations


SMARTLIST
Data-Driven
Solution

Data-Driven Solution

SMARTLIST

Specification

Machine Learning

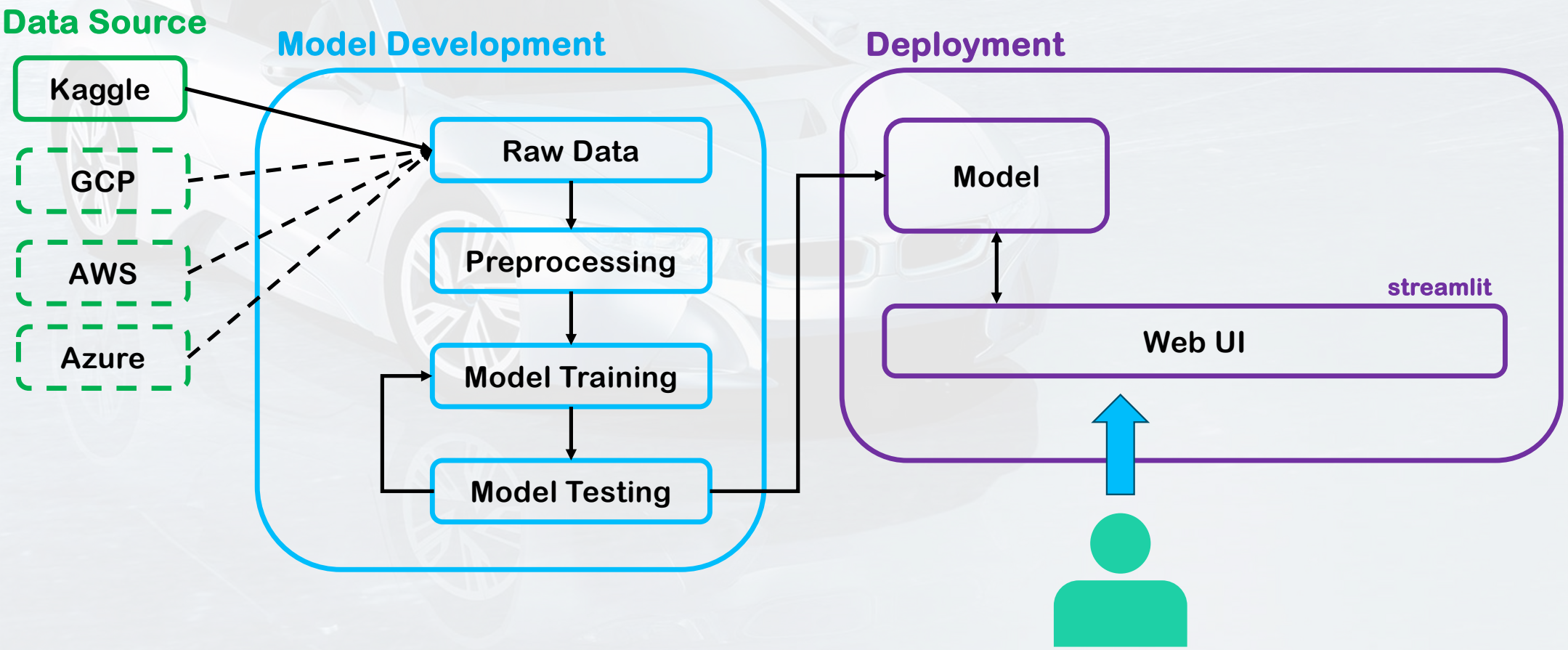
Price Range

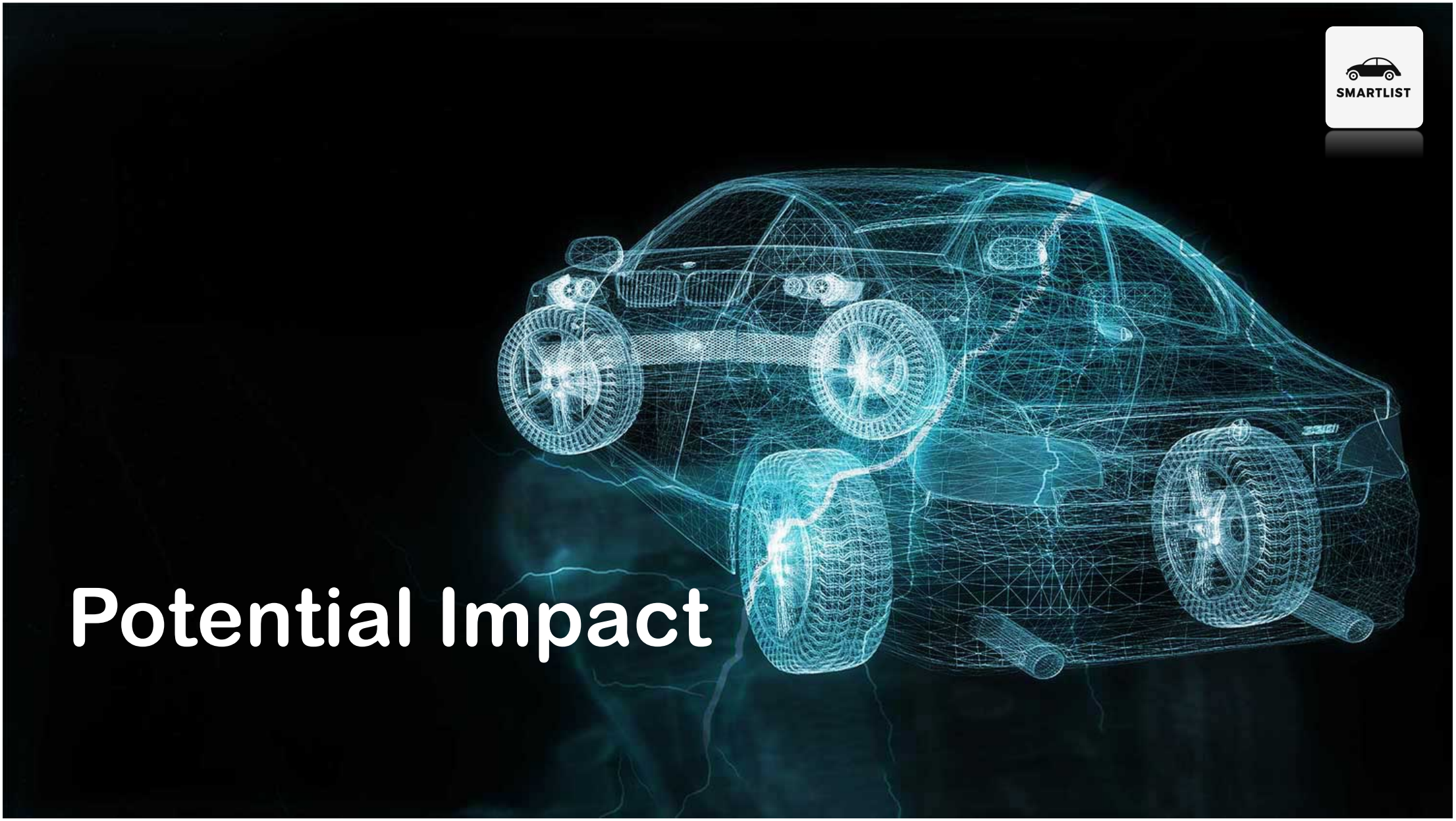
History

AI Solution

Speed Up

Architecture




SMARTLIST
Potential Impact

Industrial Impact

SMARTLIST

Speed up matching time by 25%

Market Competitors

SMARTLIST

1. eBay Motors

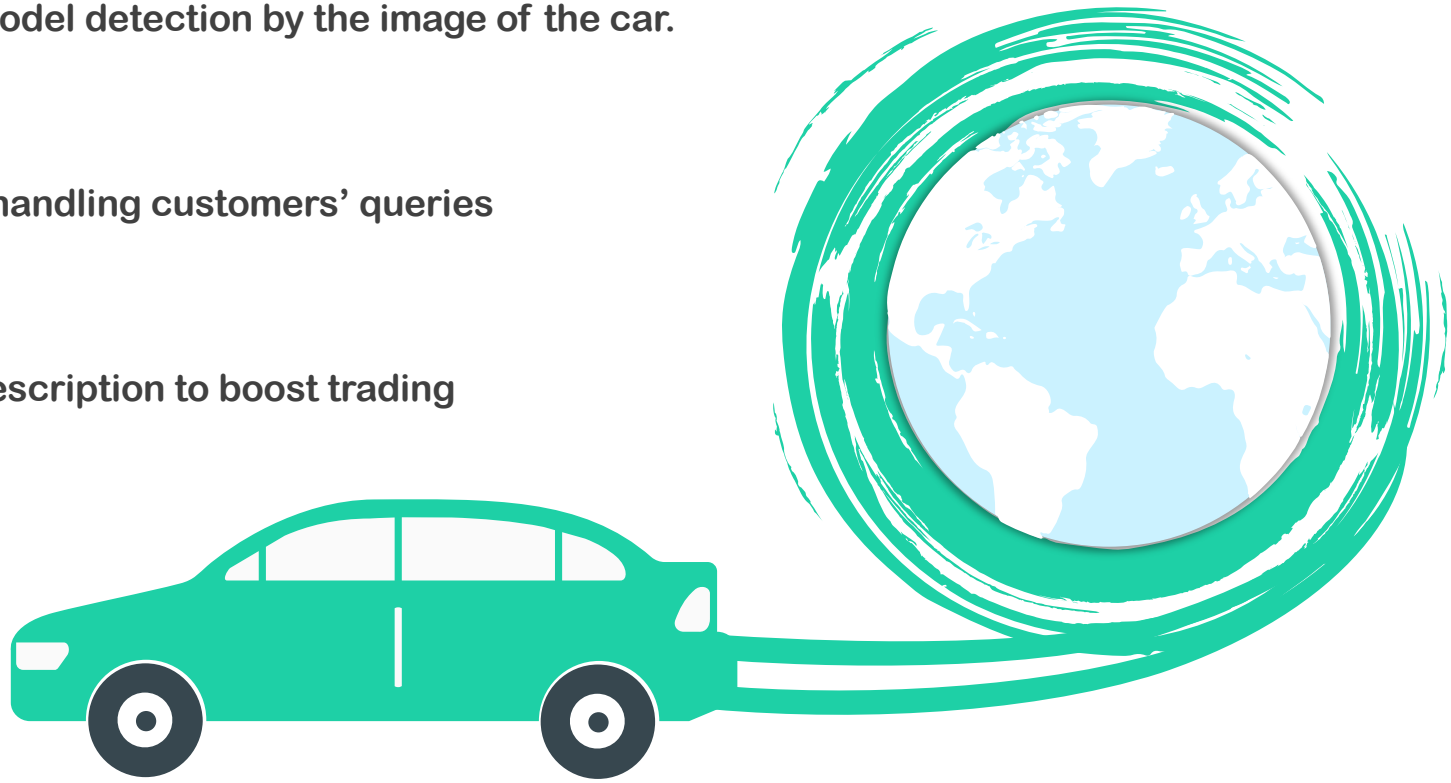
Automatic vehicle model detection by the image of the car.

2. Fullpath

Adopt ChatGPT for handling customers' queries

3. AutoRevo

Use AI-generated description to boost trading




SMARTLIST
Dataset

Dataset

SMARTLIST

Geographical

- Canada
- USA

Temporal

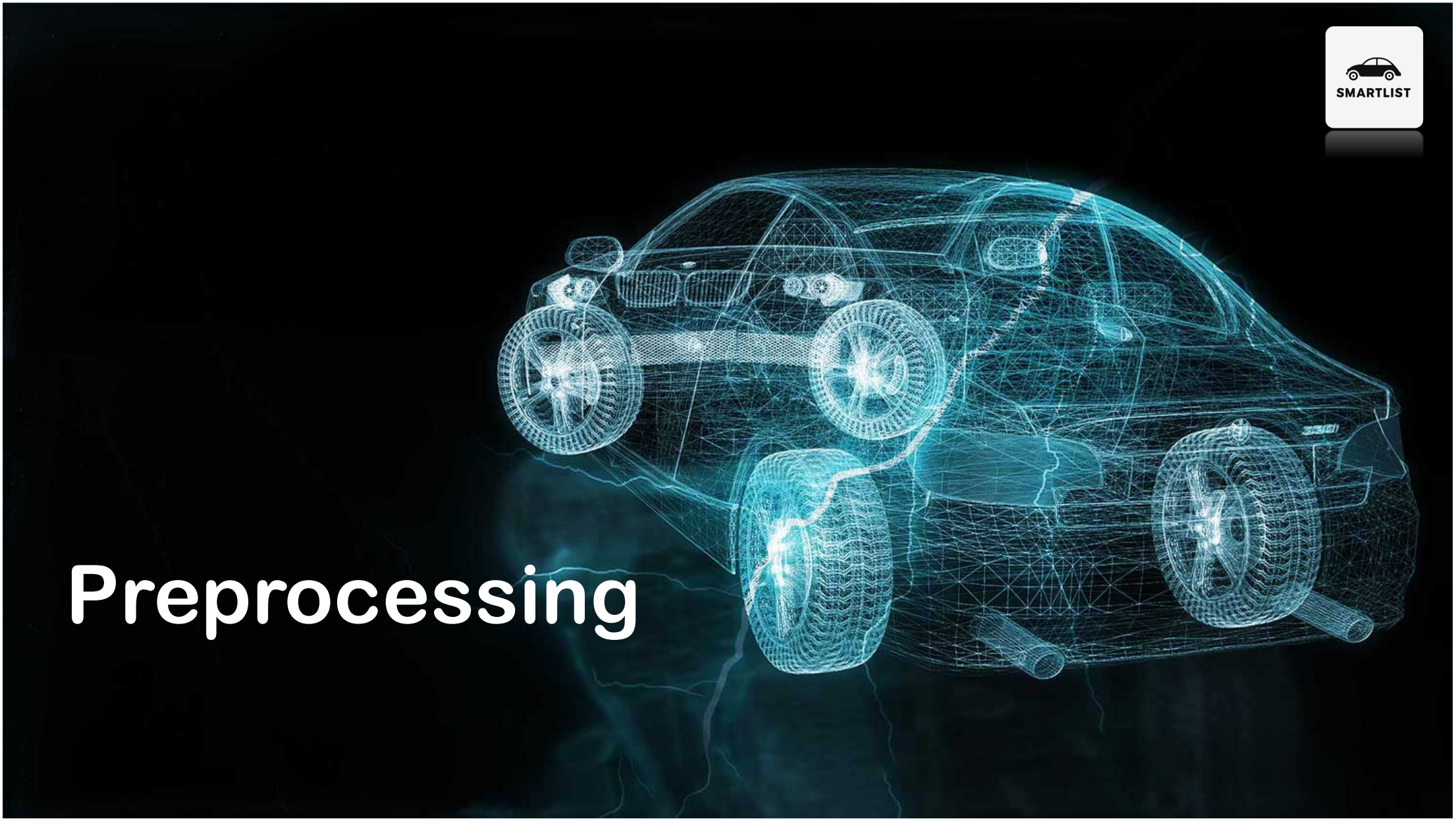
- April 2015
- June 2021

Source

- From Kaggle
- Source Link

Size

- 1.35 GB
- 7.5M Rows
  21 Columns


SMARTLIST
Preprocessing

Data Preprocessing

SMARTLIST

1 **Label Encoding**
Transform exact price to price range.

2 **Missing Value Handling**
Fill in missing value based on the occurrence probability.

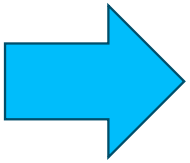
3 **Feature Flattening**
Some features are in list format which need to be transformed as vector representation.

4 **Feature Transformation**
Log transformation | One-hot Encoding | Target Encoding

Label Encoding

SMARTLIST

Price Range
> USD 60,000
USD 45,000 - 59,999
USD 30,000 - 44,999
USD 15,000 - 29,999
< USD 15,000



Price Range
5
4
3
2
1

Missing Data

SMARTLIST

Mean / Median imputation: example

Price
100
90
50
40
20
100
60
120
200

Mean = 86.66

Median = 90

Price
100
90
50
40
20
100
86.66
60
120
86.66
200

Missing Data

SMARTLIST

Make	Model	Trim
Tesla	Y	1
Tesla	Y	2
Tesla	Y	1
Tesla	X	1

Make-Model	Trim	Count	Prob
Tesla-Y	1	2	2/(2+1) = 66.67%
Tesla-Y	2	1	1/(2+1) = 33.33%
Tesla-X	1	1	1/1 = 100%

Make	Model	Trim
Tesla	Y	?
Tesla	X	?

66.67%

Make	Model	Trim
Tesla	Y	1

OR

33.33%

Make	Model	Trim
Tesla	Y	2

100%

Make	Model	Trim
Tesla	X	1

Feature Flattening

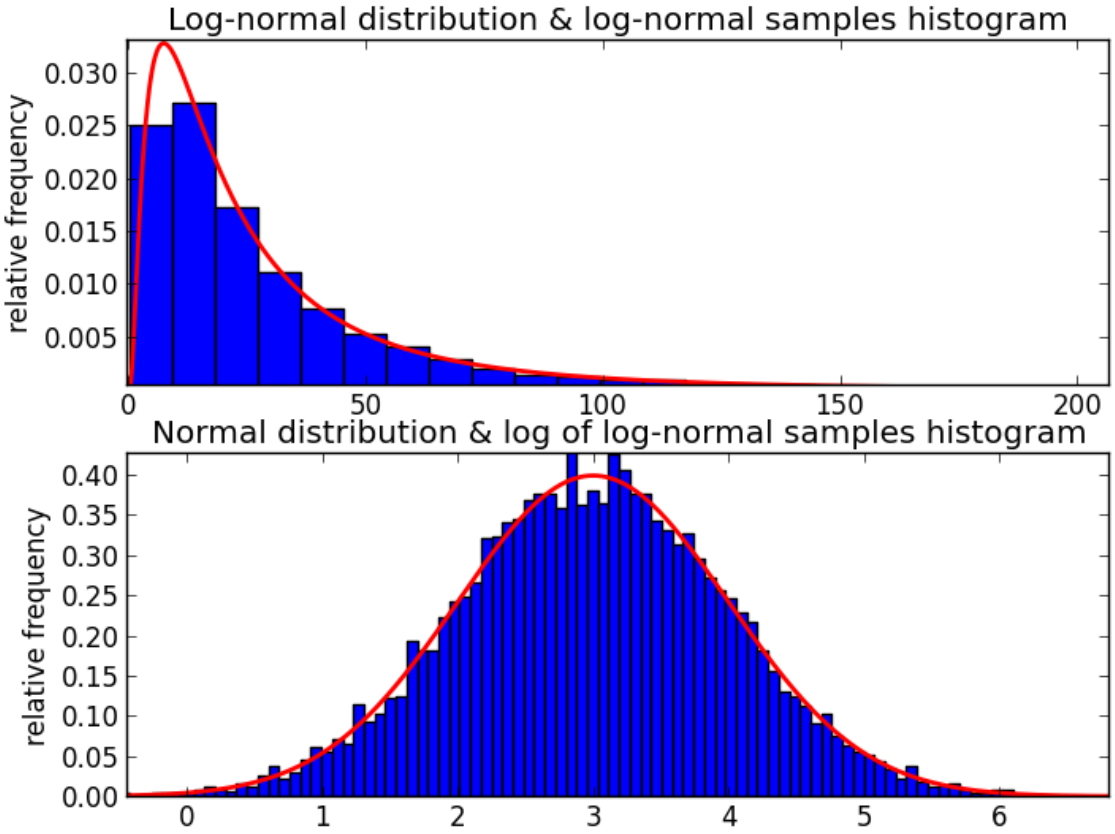
- Example:

fuel_type
Electric / Premium Unleaded

- Encoded:

Electric	Diesel	Premium Unleaded	Hydrogen
1	0	1	0

Feature Transformation




SMARTLIST
Key Insights

Business Insights

SMARTLIST

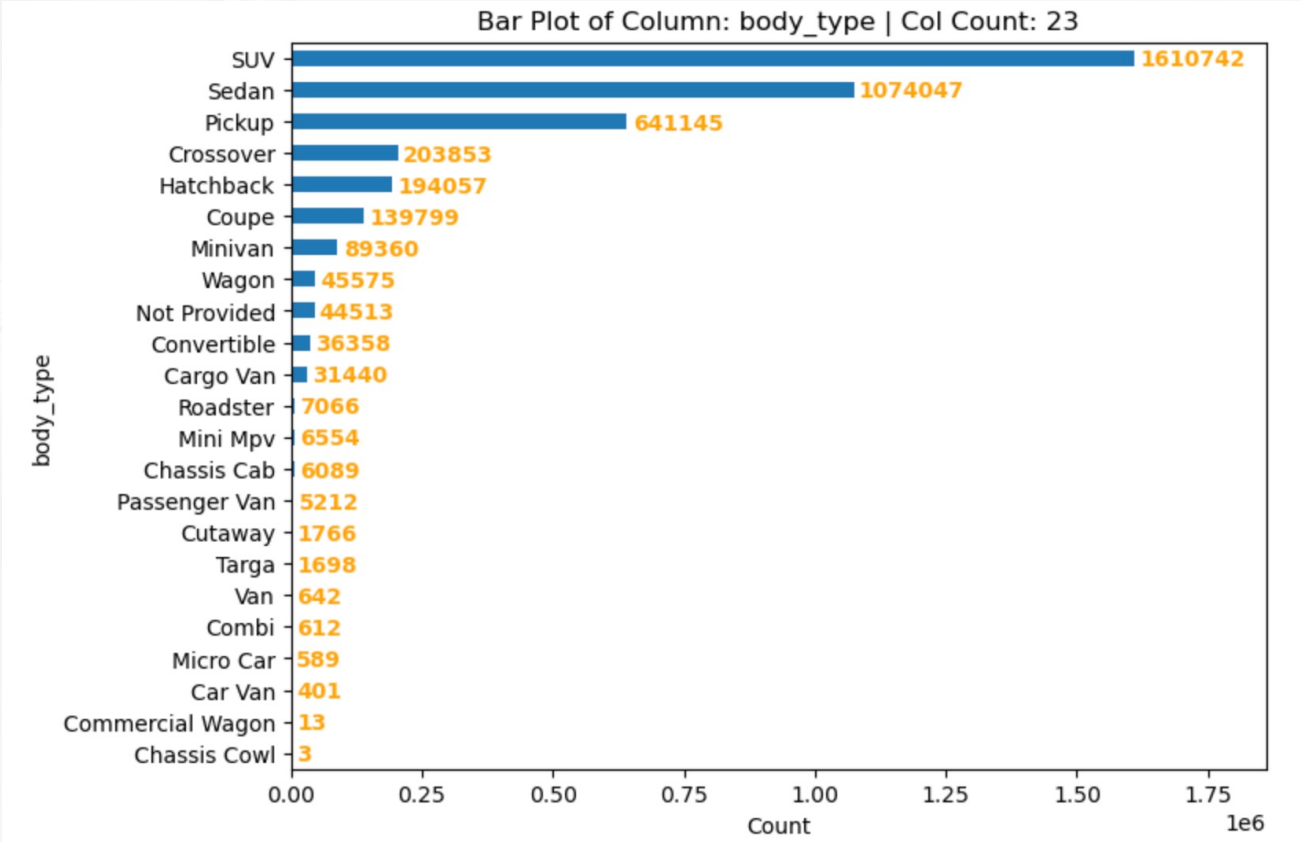
1 | **Features Analysis**
Discover pattern in different features

2 | **Feature-Price Relationship**
Explore relationship between features and pricing

Business Insights

Features Analysis - 1

- The most popular vehicle type is SUV, followed by Sedan and Pickup

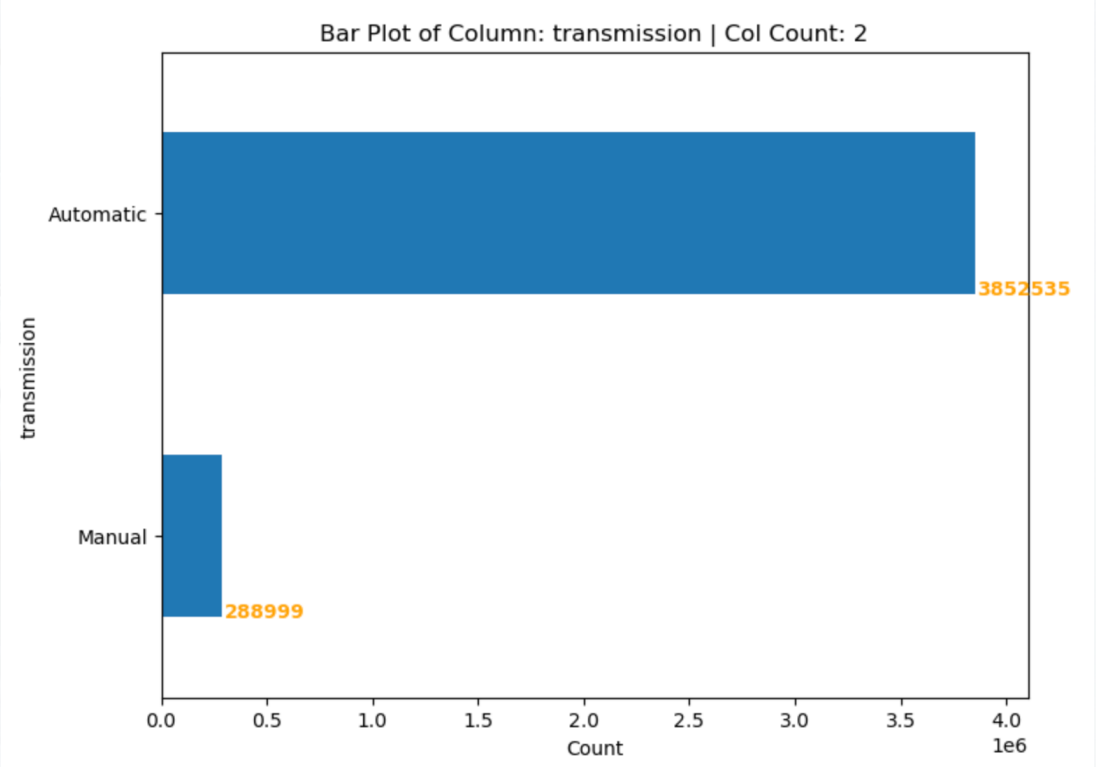


Business Insights

SMARTLIST

Features Analysis - 2

- Most of the vehicle listed are with automatic transmission

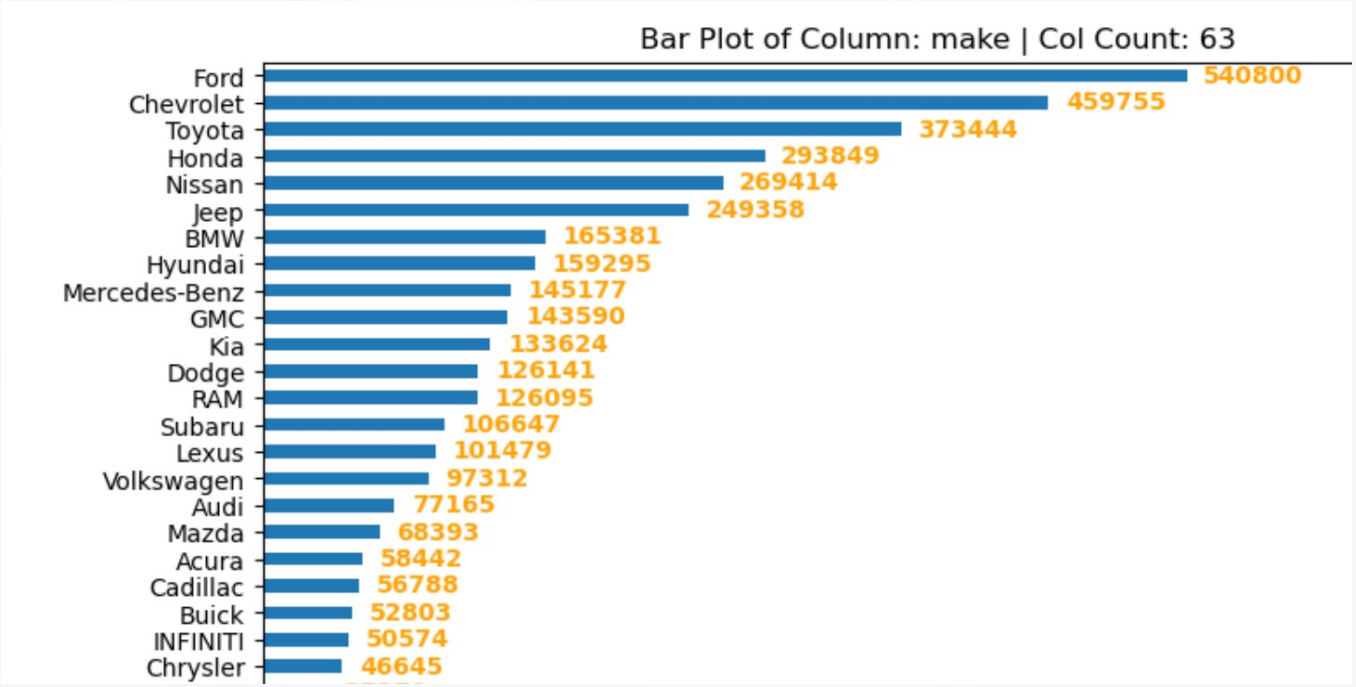


Business Insights

SMARTLIST

Features Analysis - 3

- **Top 3 popular brand:**
  1. Ford
  2. Chevrolet
  3. Toyota

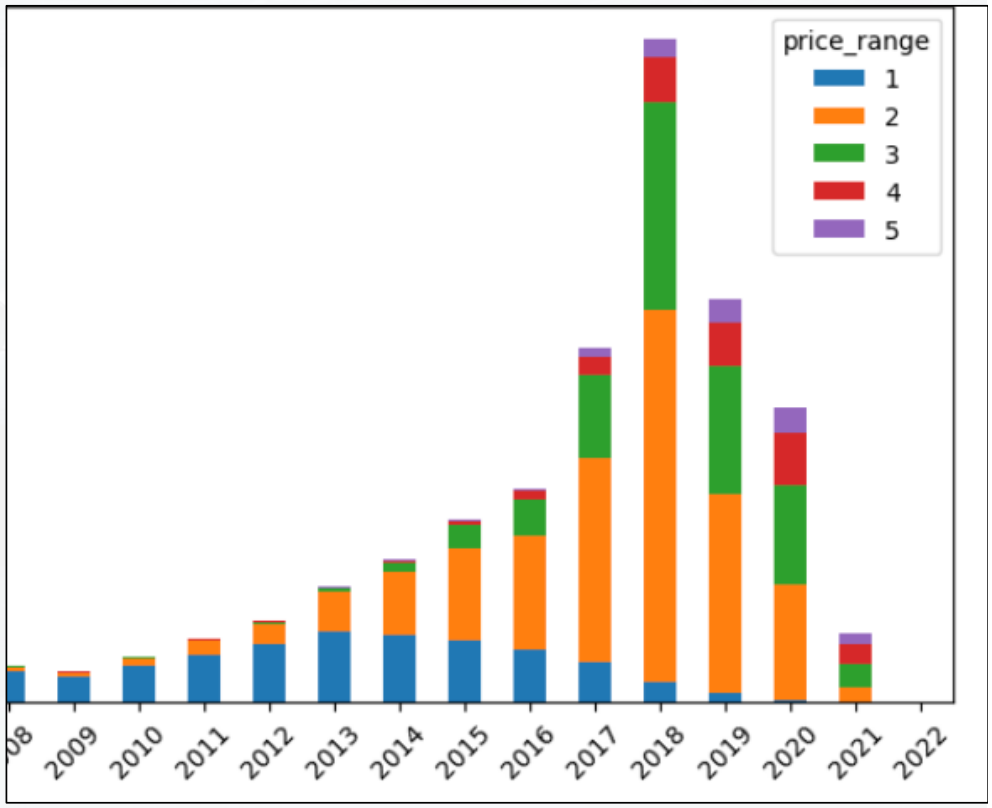


Business Insights

SMARTLIST

Feature-Pricing Relationship

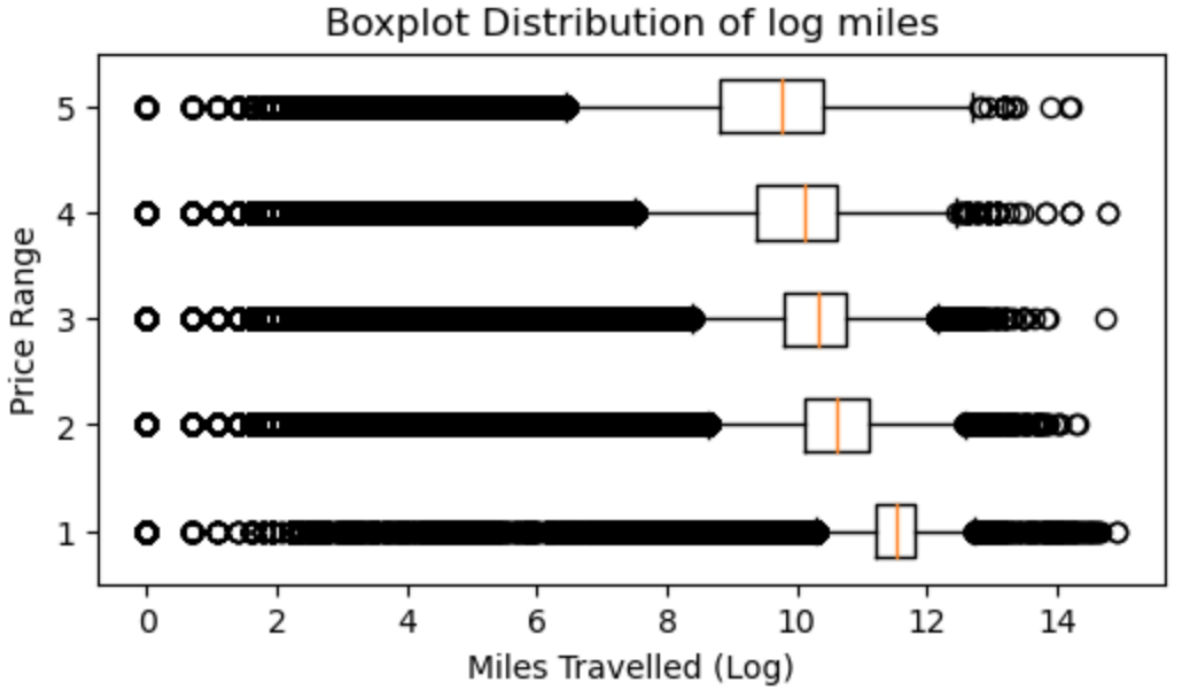
- Price vs Model Year
- For the models after 2010, the listing price start rising. It could be due to inflation



Business Insights

Feature-Pricing Relationship

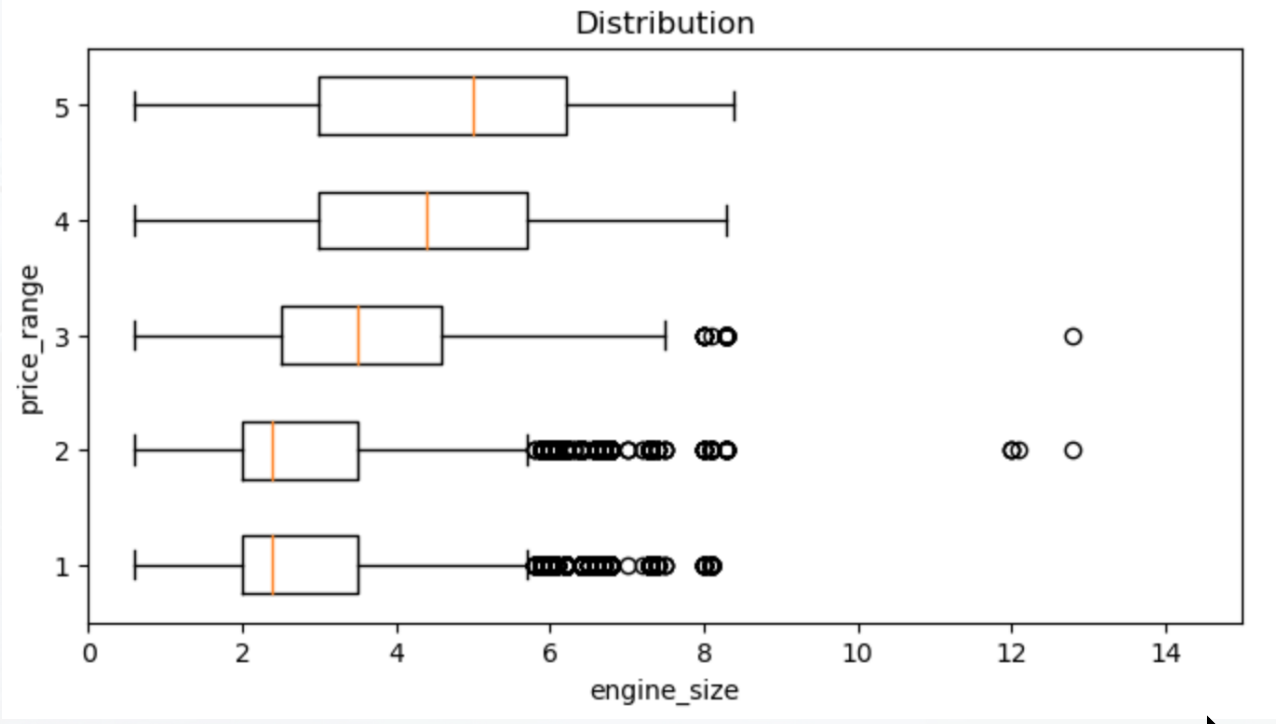
- Price vs Miles Travelled
- The more miles travelled, the less value it has.



Business Insights

Feature-Pricing Relationship

- Price vs Engine Size
- The engine size may relate to the number of cylinder in the engine.
- The more cylinder the engine has, the more expensive the car is.



Technical Insights

SMARTLIST

1 | **Log-Normal Distribution**
Apply log transformation to feature

2 | **Dimensionality Curse**
Avoid falling into dimensionality curse

3 | **Class Imbalance**
Our target variable (price range) class is skewed.

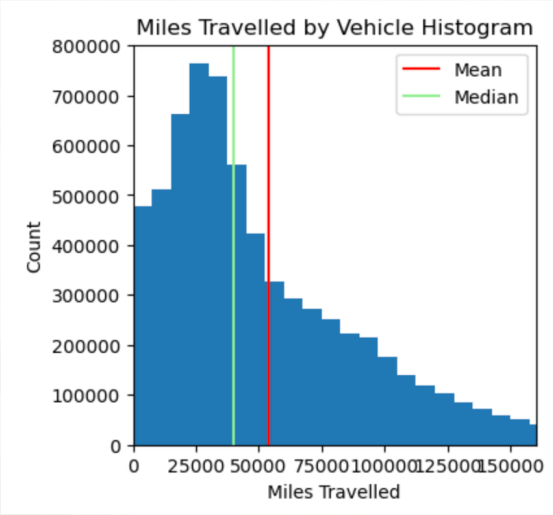
Technical Insights

SMARTLIST

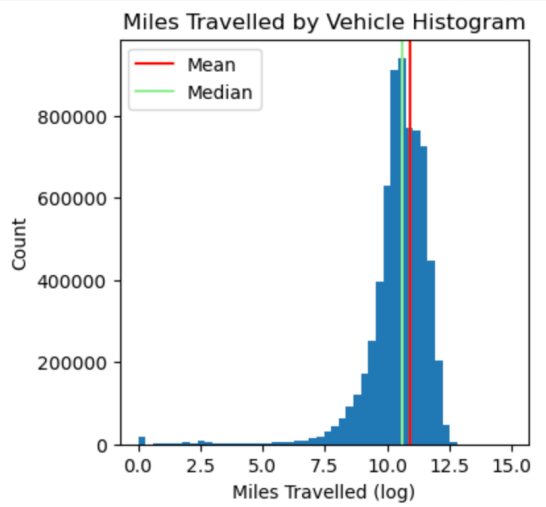
Log-Normal Distribution

- Transform the data to become “Normal Distribution”
- Many machine learning models perform better with normal distribution data / features.

Before



After

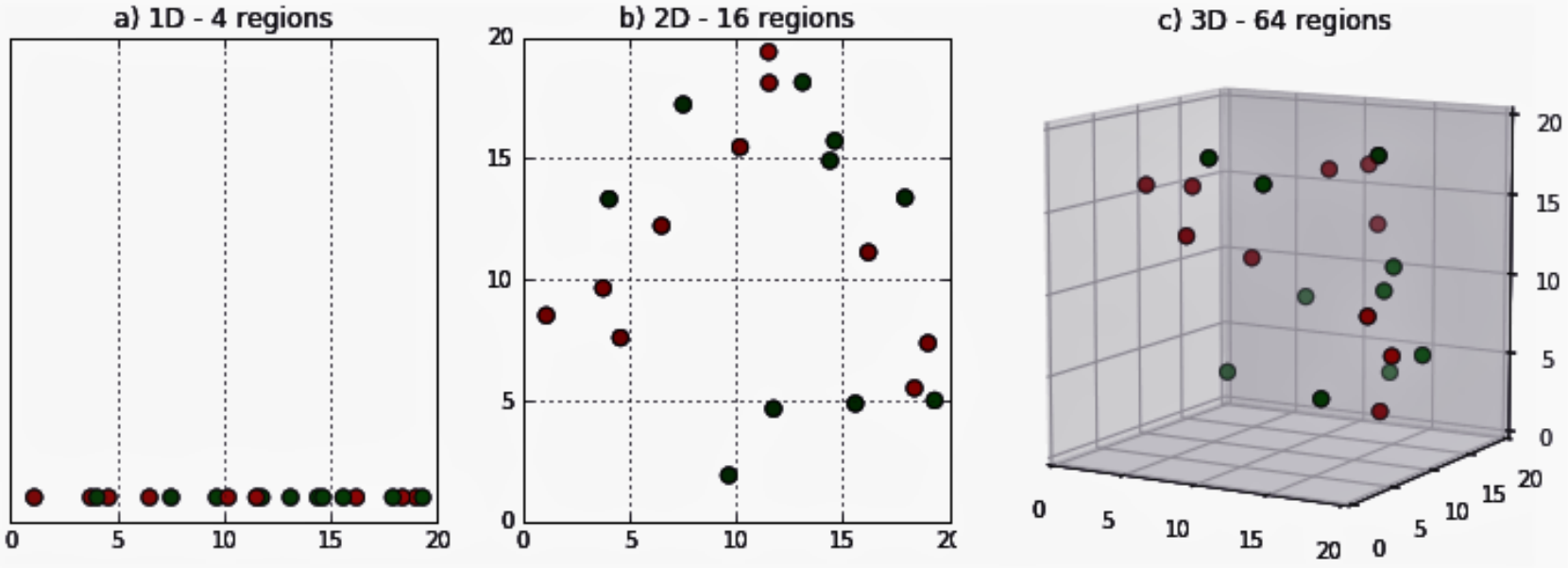


Technical Insights

SMARTLIST

Dimensionality Curse

- In data analysis, the term refers to the difficulty of finding hidden structure when the number of variables is large.

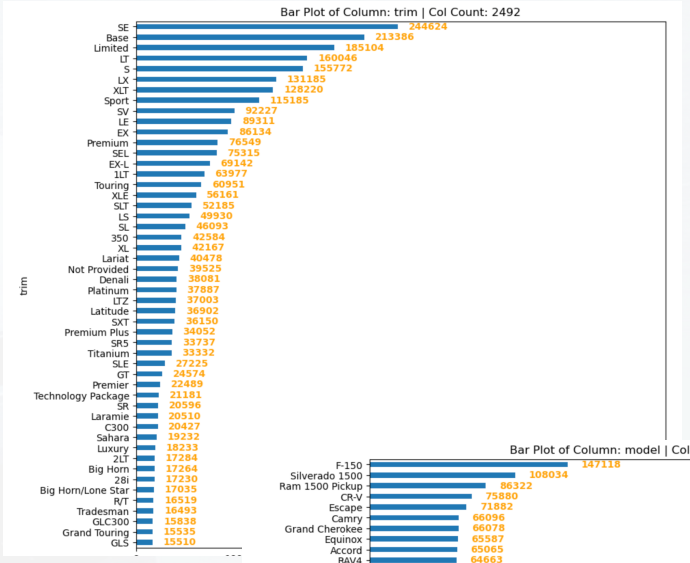


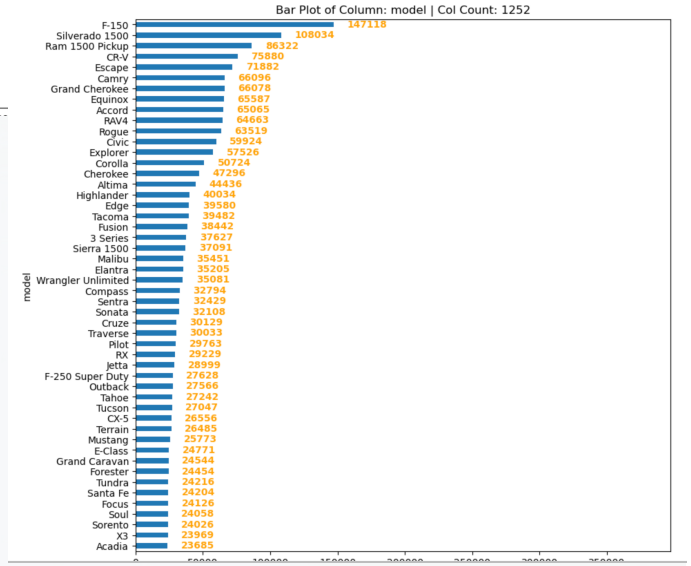
Technical Insights

SMARTLIST

Features with too many values

- **One-hot encoding**
  - ➔ Increased Dimensions
- **Count encoding**
  - ➔ Extreme Value
  - ➔ Affecting the scaling steps
- **Target encoding**
  - ➔ Does not increase dimensions
  - ➔ Encoded values between 0 ~ 1






```
Column: make has 63 distinct values
Column: model has 1252 distinct values
Column: trim has 2492 distinct values
Column: city has 6095 distinct values
Column: state has 68 distinct values
```

One-hot Encoding

SMARTLIST

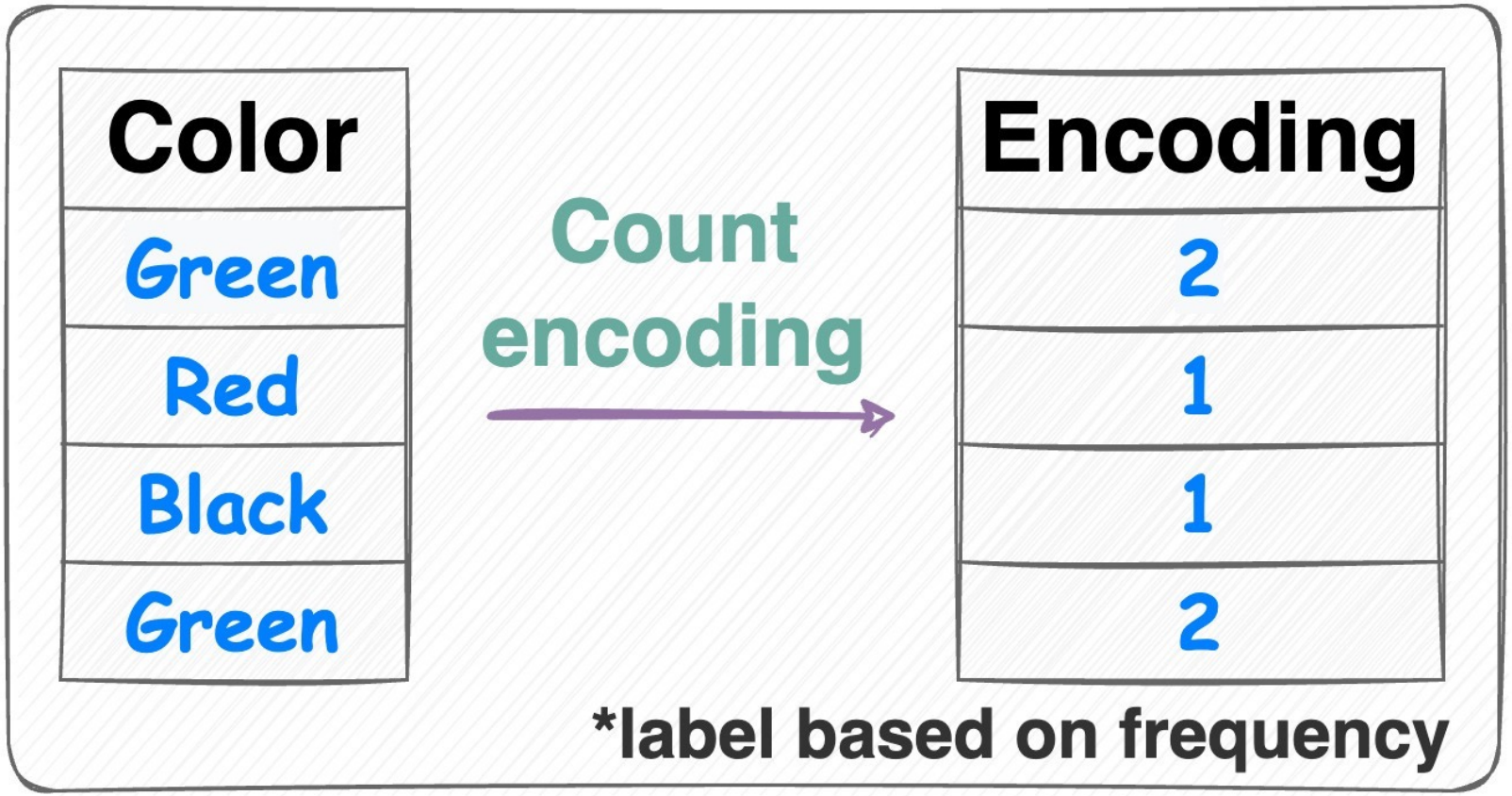
Original Data

Team	Points
A	25
A	12
B	15
B	14
B	19
B	23
C	25
C	29

One-Hot Encoded Data

Team_A	Team_B	Team_C	Points
1	0	0	25
1	0	0	12
0	1	0	15
0	1	0	14
0	1	0	19
0	1	0	23
0	0	1	25
0	0	1	29

Count Encoding



Target Encoding

Target Encoding

workclass	target
State-gov	0
Self-emp-not-inc	1
Private	0
Private	0
Private	1

workclass	target mean
State-gov	0
Self-emp-not-inc	1
Private	1/3

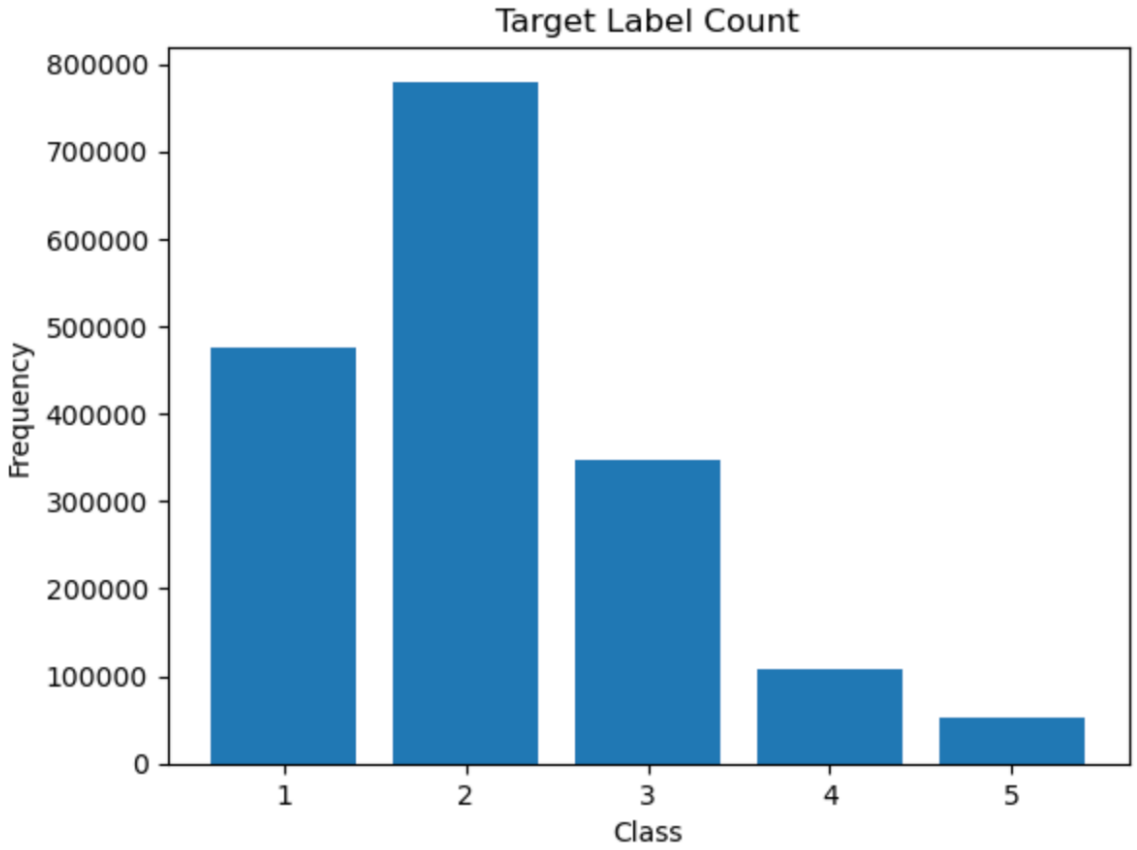
workclass
0
1
1/3
1/3
1/3

Technical Insights

Class Imbalance

- Definition:
  - Data set with skewed class proportions

- Solution:
  - Over-Sampling
  - SMOTENC
  - Hybrid-Sampling



Class Imbalance

SMARTLIST

Decision Tree

	precision	recall	f1-score	support
0	0.87	0.88	0.87	15883
1	0.85	0.86	0.86	25989
2	0.77	0.78	0.77	11588
3	0.68	0.62	0.65	3592
4	0.79	0.73	0.76	1764
accuracy			0.83	58816
macro avg	0.79	0.77	0.78	58816
weighted avg	0.83	0.83	0.83	58816

Without Sampling

	precision	recall	f1-score	support
0	0.84	0.89	0.86	15883
1	0.86	0.81	0.83	25989
2	0.74	0.76	0.75	11588
3	0.61	0.65	0.63	3592
4	0.76	0.77	0.76	1764
accuracy			0.81	58816
macro avg	0.76	0.77	0.77	58816
weighted avg	0.81	0.81	0.81	58816

With Over-Sampling

	precision	recall	f1-score	support
0	0.83	0.89	0.86	15883
1	0.87	0.81	0.84	25989
2	0.75	0.78	0.76	11588
3	0.64	0.67	0.65	3592
4	0.78	0.79	0.78	1764
accuracy			0.82	58816
macro avg	0.77	0.79	0.78	58816
weighted avg	0.82	0.82	0.82	58816

With SMOTENC-Sampling

Class Imbalance

SMARTLIST

Random Forest

	precision	recall	f1-score	support
0	0.90	0.90	0.90	15883
1	0.87	0.89	0.88	25989
2	0.80	0.82	0.81	11588
3	0.75	0.64	0.69	3592
4	0.87	0.78	0.82	1764
accuracy			0.86	58816
macro avg	0.84	0.80	0.82	58816
weighted avg	0.86	0.86	0.86	58816

Without Sampling

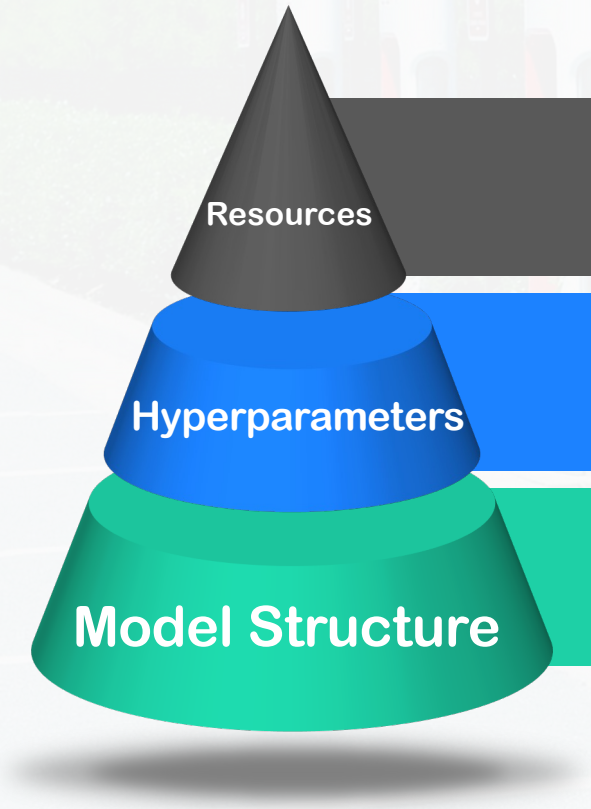
	precision	recall	f1-score	support
0	0.87	0.92	0.89	15883
1	0.90	0.84	0.87	25989
2	0.78	0.82	0.80	11588
3	0.67	0.72	0.69	3592
4	0.82	0.83	0.82	1764
accuracy			0.85	58816
macro avg	0.81	0.82	0.82	58816
weighted avg	0.85	0.85	0.85	58816

With SMOTENC - Sampling


SMARTLIST
Model Evaluation

Model Evaluation Criteria

SMARTLIST



Balance between performance and resources.

Tune the hyperparameters with selected models.

Different model structures give different results.

Model Evaluation

SMARTLIST

Model Performance (10% Data)

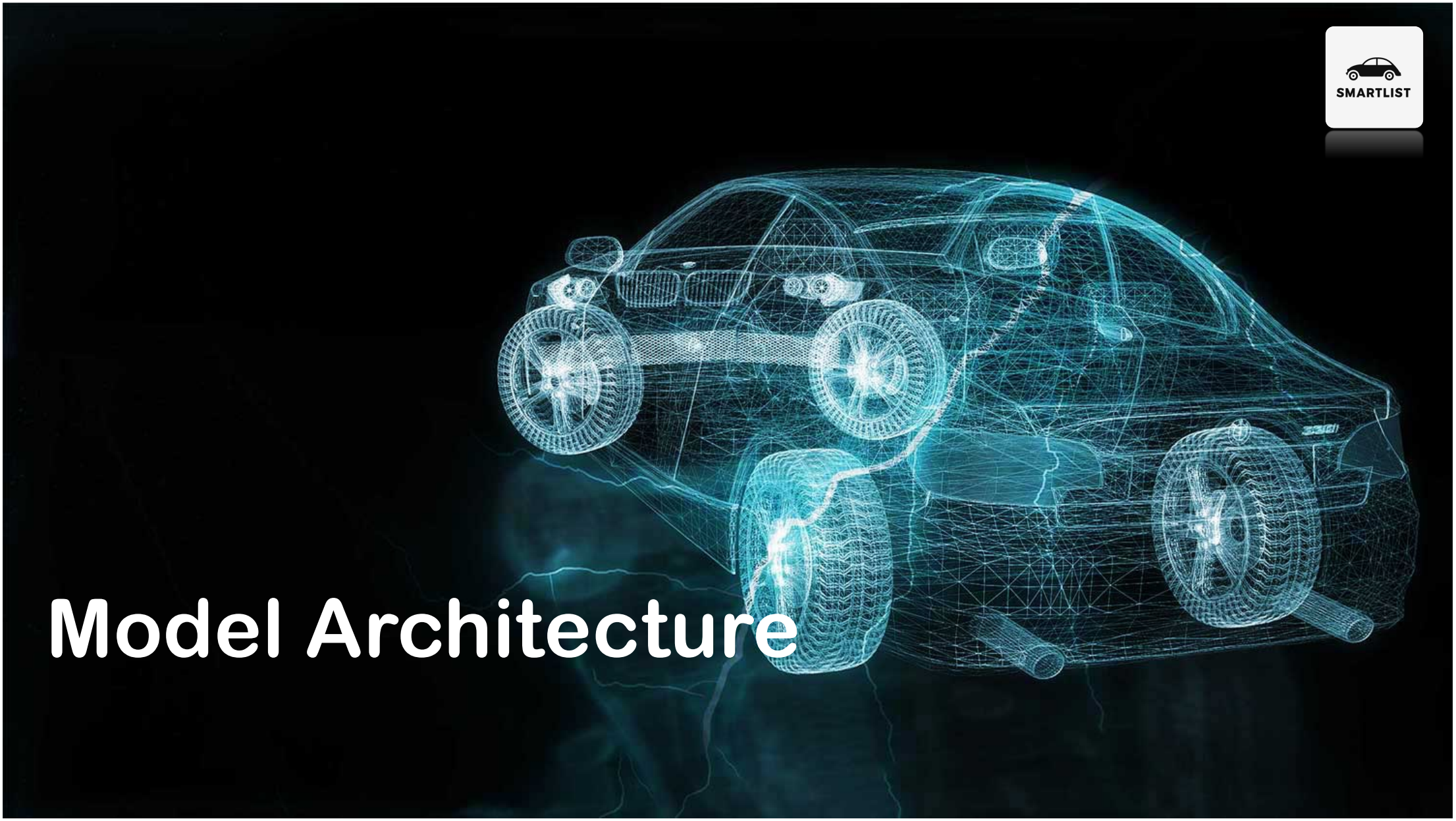
Model	Best Params / Params	Training F1 Score	Testing F1 Score	
Logistic Regression (Baseline)	{'C': 1, 'class_weight': 'balanced','max_iter': 10000,'penalty': 'l2'}	69.74%	69.82%	
Logistic Regression	{'C': 1000, 'class_weight': None, 'max_iter': 10000, 'penalty': 'l2'}	71.77%	71.81%	
Decision Tree	{'max_depth': 25, 'min_samples_split': 50}	82.15%	82.94%	
Adaptive Boosting (AdaBoost)	{'learning_rate': 1, 'n_estimators': 100}	85.34%	85.89%	✅
eXtreme Gradient Boosting (XGBoost)	{'max_depth': 15, 'n_estimators': 30}	85.42%	85.85%	✅
Random Forest (RF)	{'max_depth': 50, 'min_samples_split': 25, 'n_estimators': 150}	85.39%	85.82%	✅
Naïve Bayes	*Default*	45.75%	43.18%	
Neural Network	3 Hidden Layers (50,25,10) + ReLU Activation Layers + Epochs = 3	32.06%	32.38%	

Model Evaluation

SMARTLIST

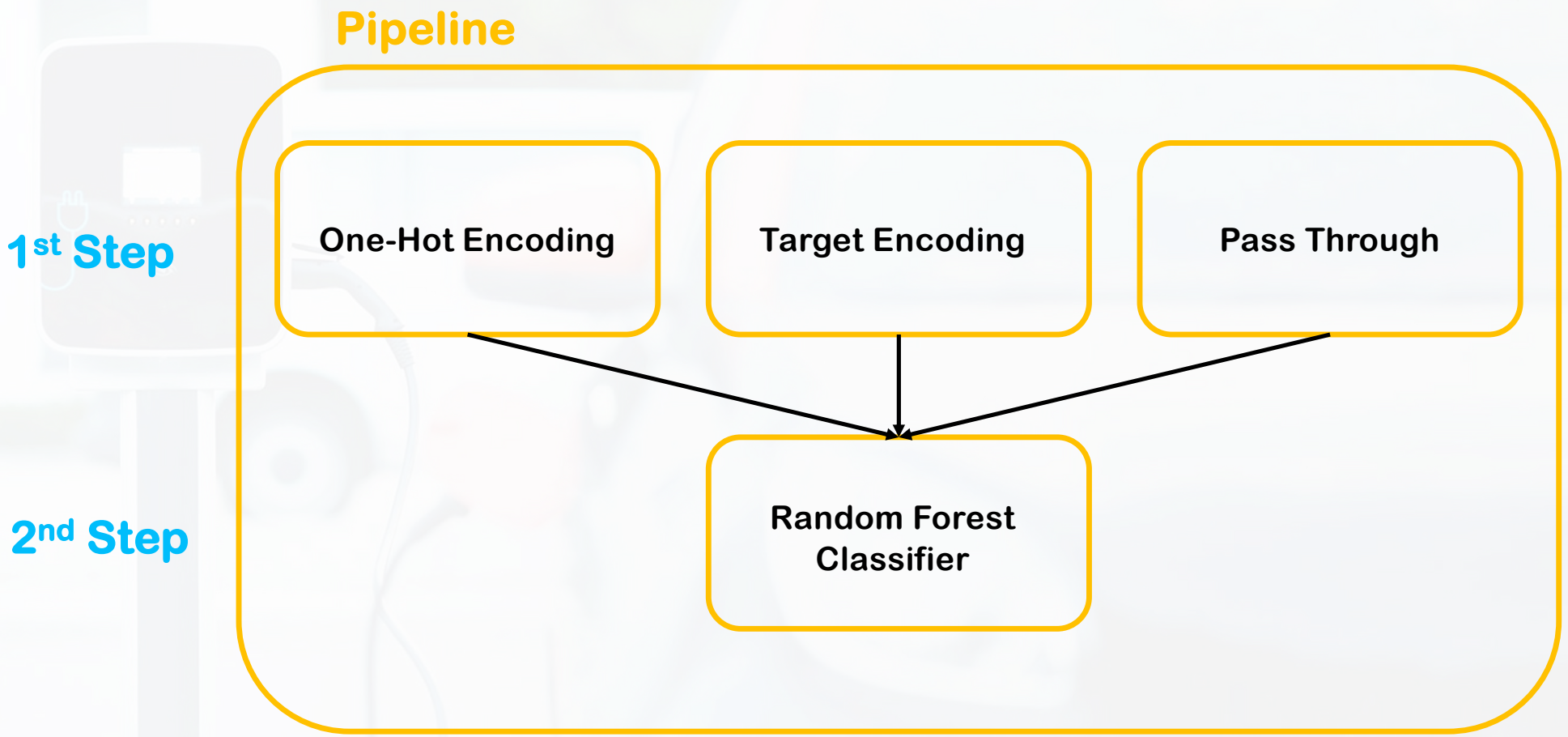
Model Performance (Full Set)

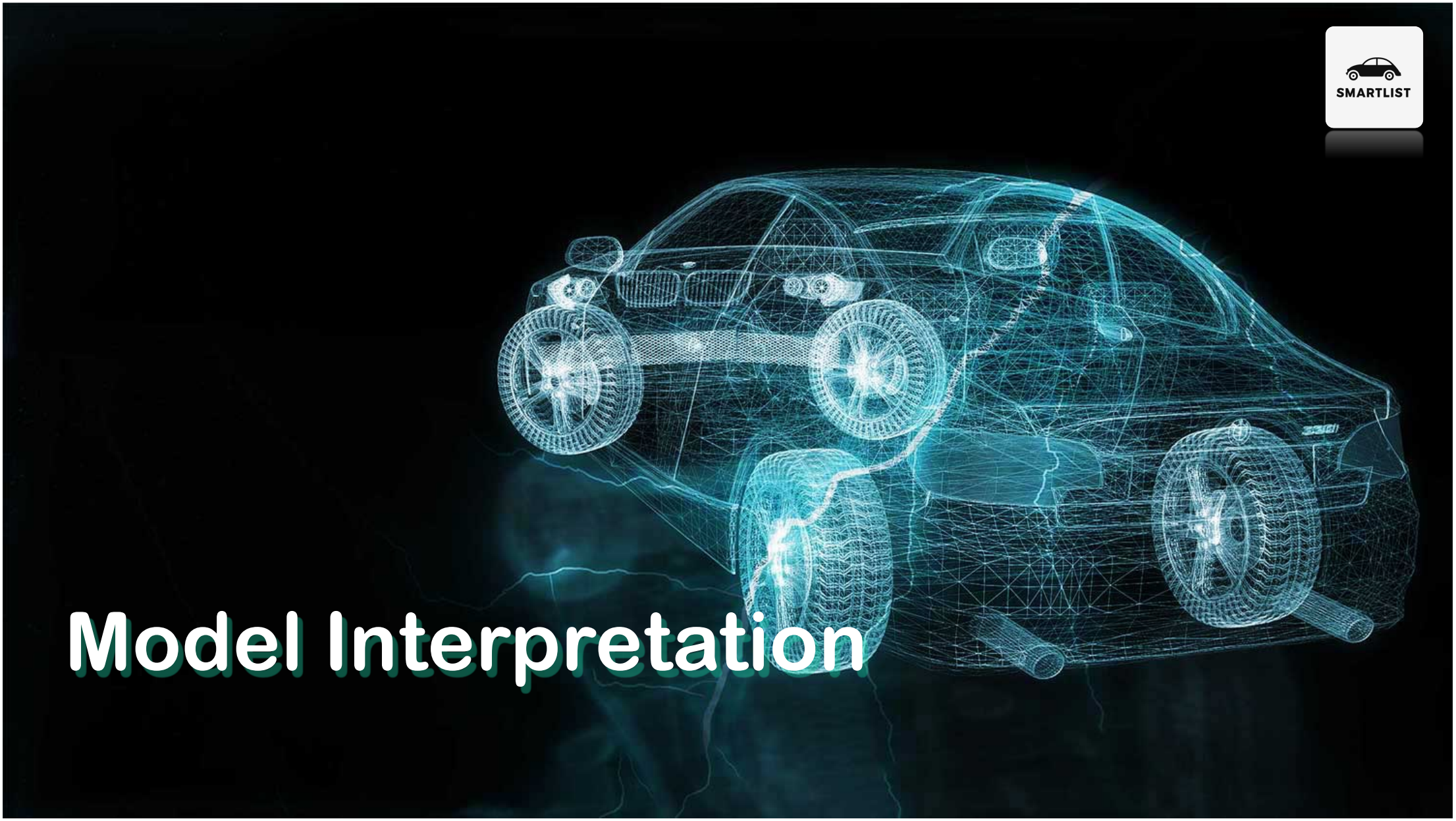
Model	Best Params / Params	Training Time	Training F1 Score	Testing F1 Score
Adaptive Boosting (AdaBoost)	{'learning_rate': 1, 'n_estimators': 100}	3h 14m 20s	86.95%	87.06%
eXtreme Gradient Boosting (XGBoost)	{'max_depth': 15, 'n_estimators': 30}	34m 43s	87.32%	87.46%
Random Forest (RF)	{'max_depth': 50, 'min_samples_split': 25, 'n_estimators': 150}	**20m 47s**	87.30%	**87.47%**


SMARTLIST
Model Architecture

Model Architecture

SMARTLIST



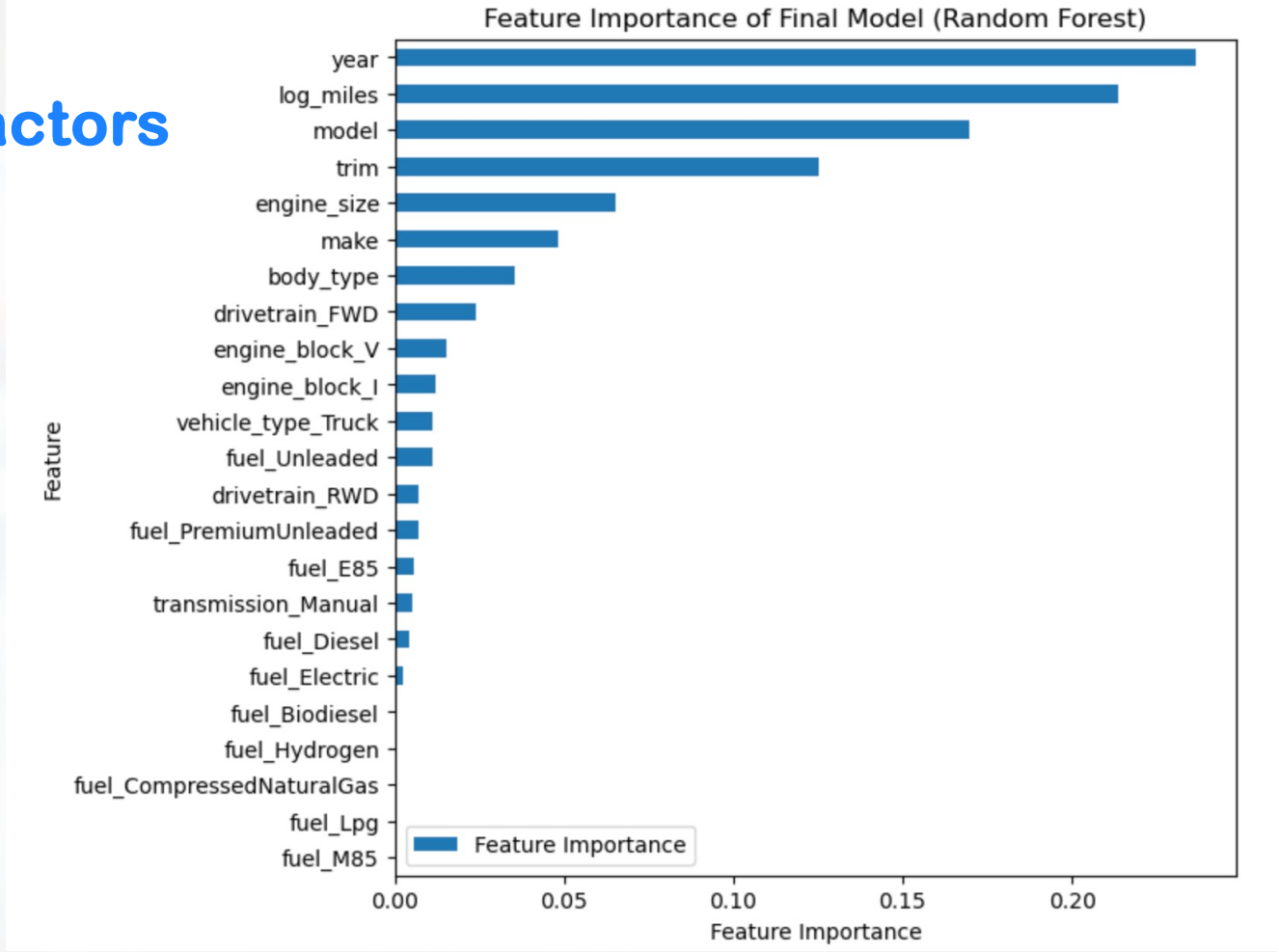

SMARTLIST
Model Interpretation

Model Interpretation

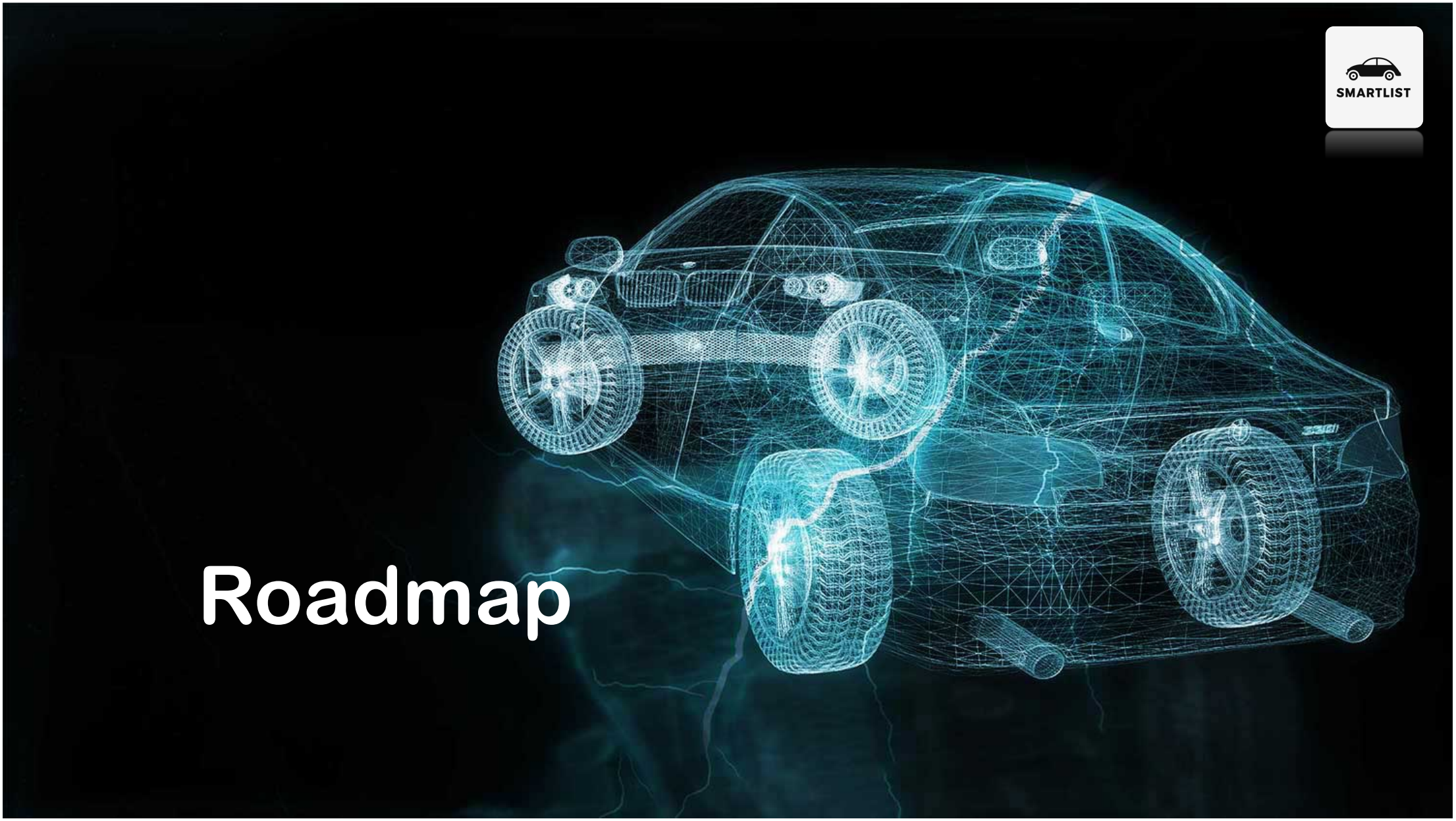
SMARTLIST

Top 5 most important factors

1. Model year
2. Miles travelled
3. Model
4. Trim (version of model)
5. Engine size



Demo Time


SMARTLIST
Roadmap

Roadmap

SMARTLIST

Sprint 0

- Problem Space Definition
- Scope Definition
- Dataset Research

Sprint 1

- Data Cleaning
- Exploratory Data Analysis
- Statistical Analysis

Sprint 2

- Data Preprocessing
- Baseline Model Training

Future

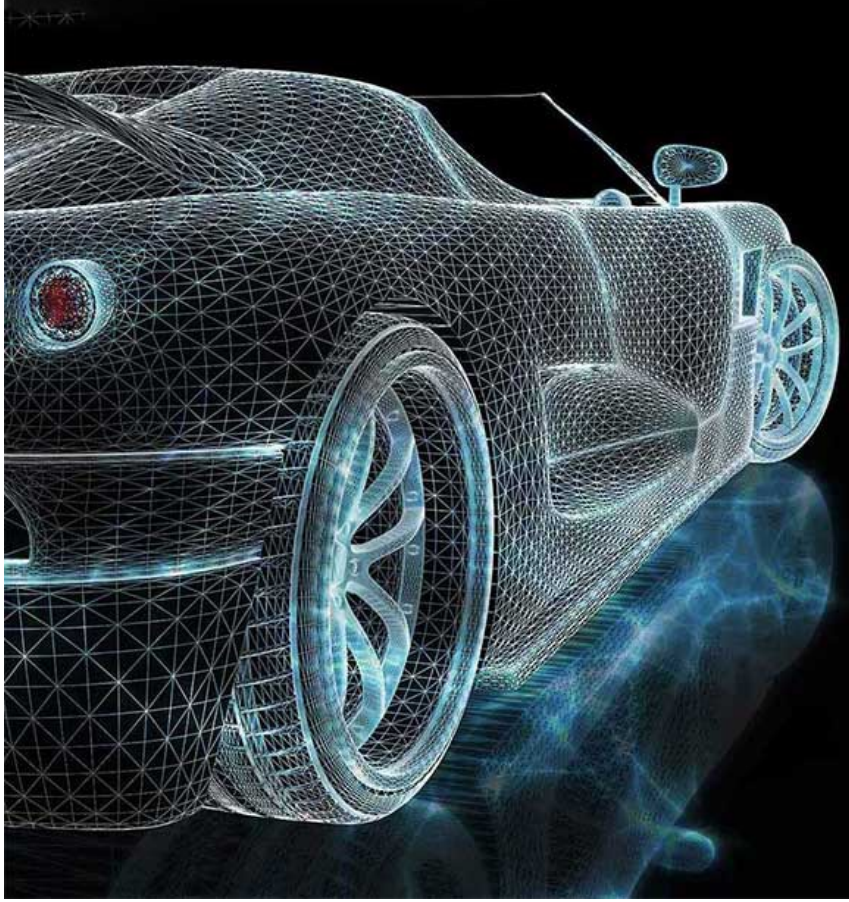
- Extend to Regression Task
- i.e. Predicting Exact Price
- Re-train model with more data
- Continually collecting new feedback
- Challenge Ourself ➔ Beat our RF model

Sprint 3

- Advance Modelling
- Web UI Integration
- Product Delivery

THANK YOU

SMARTLIST



Follow Me Now!