FOUNDATION FOR ORGANISATIONAL RESEARCH AND EDUCATION

NEW DELHI

Academic Session 2023-2025

Project-3

Customer Classification and Prediction (Car Prices) on the basis of Cluster data

Machine Learning for Managers

FMG 32 Section A

Submitted to:

Prof. Amarnath Mitra

Submitted by:

321032 – Nisha Arora

1. Project Objectives

→ The first objective is to classify the car prices data into segments or clusters using cross-validation.

→ The second objective is to classify the car prices data into segments or clusters using ensemble methods.

→ The third objective is to determine the appropriate classification model.

→ The fourth objective is to identify significant variables or features and their thresholds for classification.

2. Description of Data

2.1. Data Source, Size, Shape

2.1.1. Data Source –
https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data

2.1.2. Data Size (in KB | MB | GB …) – **88 MB**

2.1.3. Data Shape | Dimension:
Number of Variables - **16**
Number of Records – **558837**

2.2. Description of Variables

2.2.1. Index Variable(s): Car Id

2.2.2. Variables or Features having Categories | Categorical Variables or Features (CV)

2.2.2.1. Variables or Features having Nominal Categories | Categorical Variables or Features - **Nominal Type**:
make, model, trim, body, transmission, state, colour, interior, seller

2.2.2.2. Variables or Features having Ordinal Categories | Categorical Variables or Features - **Ordinal Type:** Condition

2.2.3. Non-Categorical Variables or Features: vin, odometer, mmr, selling price, sale date

Car ID: Unique identifier for each car
Year: Numeric representation of manufacturing year
Make: Brand or manufacturer of the car
Model: Specific model name of the car
Trim: Variant or version of the model
Body: Type of body style (e.g., sedan, SUV)
Transmission: Type of transmission system (e.g., automatic, manual)
VIN: Vehicle Identification Number, unique to each car
State: State where the car is located
Condition: Condition of the car, possibly ordinal categorical data
Odometer: Numeric representation of mileage
Color: Color of the car
Interior: Color or material of the interior
Seller: Entity selling the car
MMR: Market value of the car, likely non-categorical data
Selling Price: Price at which the car is sold
Sale Date: Date and time of sale

2.3. Descriptive Statistics

2.3.1. Descriptive Statistics of Outcome Categorical Variables

It provides the statistics of cluster variable (categorical variable) by giving frequency as well as relative frequency (in %).

Row ID	count	Relativ...
cluster_0	26011	23.273
cluster_1	59650	53.37
cluster_2	26106	23.358

2.3.2. Descriptive Statistics: Categorical Variables or Features
2.3.2.1. Count | Frequency Statistics

Color

Row ID	count
black	22203
white	21649
silver	16729
gray	16352
blue	10163

Model

Row ID	count
Altima	6063
F-150	2992
Fusion	2604
Camry	2460
Escape	2247

Make

Row ID	count
Ford	20837
Chevrolet	12069
Nissan	10809
Toyota	8033
Dodge	6191

Body

Row ID	count
Sedan	42596
SUV	23537
sedan	8328
suv	4968
Minivan	4348

Transmission

Row ID	count
Sedan	2
automatic	108246
manual	3514
sedan	5

2.3.3 Descriptive Statistics: Non-Categorical Variables or Features

2.3.3.1. Measures of Central Tendency

Row ID	Column	Min	Max	Mean	Std. devi...	Variance	Skewness	Kurtosis	Overall s...	No. missi...	No. NaNs	No. +∞s	No. -∞s	Median	Row count	Hi
condition	condition	1	49	30.574	13.314	177.254	-0.83	-0.197	3,417,183.716	0	0	0	0	?	111767	1
odometer	odometer	1	999,999	68,363.626	53,249.21	2,835,478,413....	1.802	12.954	7,640,797,387....	0	0	0	0	?	111767	1
mmr	mmr	25	178,000	13,782.935	9,718.146	94,442,361.104	2.026	11.693	1,540,477,346....	0	0	0	0	?	111767	25
sellingprice	sellingprice	1	171,500	13,626.721	9,787.374	95,792,682.357	1.959	10.783	1,523,017,736....	0	0	0	0	?	111767	1

2.3.3.2. Measures of Dispersion

Statistics

Rows: 4 | Columns: 12

Name	Type	# Missing val...	# Unique val...	Minimum	Maximum	25% Quantile	50% Quantile...	75% Quantile	Standard
condition	Number (dou...	0	42	1	49	24	34	41	13.314
odometer	Number (dou...	0	78138	1	999,999	28,408	52,407	99,088	53,249.21
mmr	Number (dou...	0	1066	25	178,000	7,100	12,250	18,350	9,718.146
sellingprice	Number (dou...	0	1222	1	171,500	6,900	12,100	18,250	9,787.374

Source of data-
https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data

3. Analysis of Data
3.1. Data Pre-Processing
3.1.1. Missing Data Statistics and Treatment

3.1.1.1.1. Missing Data Statistics: 16

3.1.1.1.2. Missing Data Treatment: make, model, trim, body, transmission, state, colour, interior, seller, condition, vin, odometer, mmr, selling price, sale date

3.1.1.1.2.1. Removal of Records with More Than 50% Missing Data

3.1.1.2.1. Missing Data Statistics: Categorical Variables or Features

Name	# Missing values
year	0
make	2141
model	2170
trim	2203
body	2688
transmission	13241
state	0
color	163
interior	163
seller	0

3.1.1.2.2. Missing Data Treatment: Categorical Variables or Features - 10

3.1.1.2.2.1. Removal of Variables or Features with More Than 50% Missing Data:

make, model, trim, body, transmission, state, colour, interior, seller, condition

3.1.1.2.2.2. Imputation of Missing Data using Descriptive Statistics: Mode

3.1.1.3.1. Missing Data Statistics: Non-Categorical Variables or Features

Name	# Missing values
vin	2
condition	2342
odometer	21
mmr	9
sellingprice	2
saledate	2

3.1.1.3.2. Missing Data Treatment: Non-Categorical Variables or Features - 6

3.1.1.3.2.1. Removal of Variables or Features with More Than 50% Missing Data: vin, odometer, mmr, selling price, sale date

3.1.1.3.2.2. Imputation of Missing Data using Descriptive Statistics: Mean

3.1.2. Numerical Encoding of Categorical Variables or Features (Encoding Schema - Alphanumeric Order)

- In this case, category to number node will be used to encode the categorical variables.

Color-
8 – black
9 – blue
14 – gray
22 – silver
24 – white

Model
30-Altima
91- F-150
 90- Fusion
62- Camry
75- Escape

Make
 19-Ford
0-Chevrolet
5-Nissan
17-Toyota
22-Dodge

Body
0- Sedan
1- SUV
28-sedan
44-suv
9-Minivan

Transmission
0 – Sedan
1- Automatic
2 – Manual
3 – sedan

3.1.3. Outlier Statistics and Treatment (Scaling | Transformation)

3.1.3.1.1. Outlier Statistics: Non-Categorical Variables or Features

Row ID	S Outlier ...	I Membe...	I Outlier ...	D Lower ...	D Upper ...
Row0	condition	111767	0	-1.5	66.5
Row1	odometer	111767	2066	-77,611	205,105
Row2	mmr	111767	3244	-9,775	35,225
Row3	sellingprice	111767	3222	-10,125	35,275

3.1.3.1.2. Outlier Treatment: Non-Categorical Variables or Features
3.1.3.1.2.1. Standardization
3.1.3.1.2.2. Normalization using Min-Max Scaler:

Min-max normalization, also known as feature scaling, is a technique

used in data preprocessing to scale numerical features to a specific range, typically between 0 and 1.
The formula for min-max normalization is:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

3.1.3.1.2.3. Log Transformation

3.1.4. Data Bifurcation: Training & Testing Sets

The training and testing data have been bifurcated into 70% and 30% respectively.

3.2. Data Analysis

3.2.1. Cross-Validation using Decision Tree

Cross-validation using a decision tree involves splitting the dataset into k subsets, training the decision tree on k-1 subsets and validating on the remaining subset by repeating this process k times and averaging the results to assess the model's performance and generalization ability.

3.2.2. Cross-Validation using Other Methods

3.2.2.1.Logistic Regression

Cross-validation with logistic regression involves partitioning the dataset into training and validation sets, fitting the logistic regression model on the training data and evaluating its performance on the validation set. This process is repeated multiple times with different partitions to estimate the model's generalization performance and minimize overfitting.

3.2.2.2. K-Nearest Neighbours

Cross-validation with KNN entails splitting the dataset into training and validation sets, then iterating through different values of k (number of nearest neighbours) to find the optimal k value that minimizes error on the validation set. This process helps assess the KNN model's performance and its ability to generalize to new data.

3.2.3. Ensemble Method using Random Forest

Random forest is an ensemble learning method where multiple decision trees are trained on random subsets of the data and features. During prediction, each tree votes on the outcome and the final prediction is determined by the majority vote. This approach improves prediction accuracy and reduces overfitting compared to individual decision trees.

3.2.4. Ensemble Method using XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm that uses a gradient boosting framework. It sequentially builds multiple decision trees, each correcting the errors of the previous one. XGBoost incorporates regularization techniques to prevent overfitting and is known for its efficiency and effectiveness in various machine learning tasks.

3.2.1.1. Model Performance Evaluation of Cross-Validation using Decision Tree

Without Pruning

Row ID	cluster_1	cluster_0	cluster_2
cluster_1	17895	0	0
cluster_0	0	7804	0
cluster_2	0	0	7832

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_1	17895	0	15636	0	1	1	1	1	1	?	?
cluster_0	7804	0	25727	0	1	1	1	1	1	?	?
cluster_2	7832	0	25699	0	1	1	1	1	1	?	?
Overall	?	?	?	?	?	?	?	?	?	1	1

With Pruning

Row ID	cluster_1	cluster_0	cluster_2
cluster_1	17895	0	0
cluster_0	0	7804	0
cluster_2	0	0	7832

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_1	17895	0	15636	0	1	1	1	1	1	?	?
cluster_0	7804	0	25727	0	1	1	1	1	1	?	?
cluster_2	7832	0	25699	0	1	1	1	1	1	?	?
Overall	?	?	?	?	?	?	?	?	?	1	1

Cluster 0

• This cluster has a high number of true positives and true negatives indicating that the model correctly classified most instances within this cluster.

• The precision and recall scores are both very high suggesting that the model effectively identifies true positives while also minimizing false positives.

Cluster 1

• This cluster has a lower recall and precision compared to cluster 0, indicating that the model's performance is not as strong for this segment.

• The number of false positives is relatively high, suggesting that the model may misclassify some instances within this cluster.

• Despite the lower performance metrics, the specificity is very high indicating that the model correctly identifies true negatives within this cluster.

Cluster 2

• This cluster has a relatively high recall and precision, indicating that the model performs well.

• The number of false positives is relatively low suggesting that the model effectively minimizes misclassifications within this cluster.

• Both sensitivity and specificity scores are high indicating that the model correctly identifies both true positives and true negatives within this cluster.

Comparative analysis of decision tree with and without pruning

Pruning generally improves precision and specificity while slightly reducing recall and sensitivity. Pruning removes unnecessary branches from the tree, simplifying the model and reducing overfitting. This can lead to better generalization and potentially improved performance on unseen data.

We didn't observe a significant difference between a pruned and non-pruned decision tree in our case. It may be because:

1. **Dataset characteristics:** The data we used might be relatively simple, and the decision tree without pruning may not have overfit considerably.
2. **Pruning settings:** The pruning settings in KNIME's Decision Tree Learner node might have been configured in a way that resulted in minimal removal of branches.
3. **Randomness:** There can be an element of randomness in decision tree generation. Rerunning the experiment with both pruned and non-pruned trees might yield a slight difference on another iteration.

The choice of whether to prune the decision tree depends on the specific requirements of the problem and the trade-off between precision and recall. If minimizing false positives is crucial (can be used for risk assessment) pruning may be preferred. If capturing as many true positives as possible is more important (can be used for customer retention) pruning may be avoided.

3.2.2.1. Model Performance Evaluation of Cross-Validation using Other Methods

Logistic Regression

Row ID	Logit	Variable	
Row1	cluster_0	year	
Row2	cluster_0	transmission=automatic	
Row3	cluster_0	transmission=manual	
Row4	cluster_0	transmission=sedan	
Row5	cluster_0	state=3vwd17aj4fm236636	
Row6	cluster_0	state=3vwd17aj5fm219943	
Row7	cluster_0	state=3vwd17aj5fm221322	
Row8	cluster_0	state=3vwd17aj5fm225953	
Row9	cluster_0	state=3vwd17aj7fm222388	
Row10	cluster_0	state=3vwd17aj7fm229552	
Row11	cluster_0	state=ab	
Row12	cluster_0	state=al	
Row13	cluster_0	state=az	
Row14	cluster_0	state=ca	
Row15	cluster_0	state=co	
Row90	cluster_0	color=silver	-0.003
Row91	cluster_0	color=turquoise	0.052
Row92	cluster_0	color=white	0.3
Row93	cluster_0	color=yellow	0.219
Row94	cluster_0	color=â€"	-0.368
Row95	cluster_0	interior=black	-0.159
Row96	cluster_0	interior=blue	0.19
Row97	cluster_0	interior=brown	-0.292
Row98	cluster_0	interior=burgundy	0.009
Row99	cluster_0	interior=gold	-0.112
Row100	cluster_0	interior=gray	-0.041
Row101	cluster_0	interior=green	-0.074
Row102	cluster_0	interior=off-white	-0.087
Row103	cluster_0	interior=orange	0.033
Row104	cluster_0	interior=purple	0.076
Row105	cluster_0	interior=red	0.082
Row106	cluster_0	interior=silver	0.169
Row107	cluster_0	interior=tan	-0.047
Row108	cluster_0	interior=white	0.033
Row109	cluster_0	interior=yellow	0.001
Row110	cluster_0	interior=â€"	-0.269
Row111	cluster_0	Car id (to number)	-0
Row112	cluster_0	make (to number)	-0.001
Row113	cluster_0	model (to number)	0
Row114	cluster_0	trim (to number)	0
Row115	cluster_0	body (to number)	-0.032
Row116	cluster_0	transmission (to number)	-0.076
Row117	cluster_0	state (to number)	-0.02
Row118	cluster_0	color (to number)	0.049
Row119	cluster_0	interior (to number)	-0.031
Row120	cluster_0	seller (to number)	0
Row121	cluster_0	Clusters (to number)	-72.932
Row122	cluster_0	odometer	0.625
Row123	cluster_0	mmr	-0.346
Row124	cluster_0	sellingprice	0.312

Row215	cluster_1	color=silver	-0.018
Row216	cluster_1	color=turquoise	-0.013
Row217	cluster_1	color=white	0.694
Row218	cluster_1	color=yellow	0.289
Row219	cluster_1	color=â€”	-1.035
Row220	cluster_1	interior=black	-0.671
Row221	cluster_1	interior=blue	0.432
Row222	cluster_1	interior=brown	-0.99
Row223	cluster_1	interior=burgundy	-0.04
Row224	cluster_1	interior=gold	0.107
Row225	cluster_1	interior=gray	-0.546
Row226	cluster_1	interior=green	-0.007
Row227	cluster_1	interior=off-white	-0.055
Row228	cluster_1	interior=orange	-0.119
Row229	cluster_1	interior=purple	0.06
Row230	cluster_1	interior=red	0.147
Row231	cluster_1	interior=silver	-0.093
Row232	cluster_1	interior=tan	0.213
Row233	cluster_1	interior=white	0.246
Row234	cluster_1	interior=yellow	0.032
Row235	cluster_1	interior=â€”	-0.645
Row236	cluster_1	Car id (to number)	-0
Row237	cluster_1	make (to number)	-0.015
Row238	cluster_1	model (to number)	-0
Row239	cluster_1	trim (to number)	0
Row240	cluster_1	body (to number)	0.058
Row241	cluster_1	transmission (to number)	-0.187
Row242	cluster_1	state (to number)	0.008
Row243	cluster_1	color (to number)	0.152
Row244	cluster_1	interior (to number)	-0.13
Row245	cluster_1	seller (to number)	0
Row246	cluster_1	Clusters (to number)	-149.664
Row247	cluster_1	odometer	0.943
Row248	cluster_1	mmr	-0.192
Row249	cluster_1	sellingprice	-0.21

Row141	cluster_1	state=fl	-0.138
Row142	cluster_1	state=ga	-0.067
Row143	cluster_1	state=hi	-0.046
Row144	cluster_1	state=il	0.104
Row145	cluster_1	state=in	0.474
Row146	cluster_1	state=la	0.234
Row147	cluster_1	state=ma	0.046
Row148	cluster_1	state=md	-0.619
Row149	cluster_1	state=mi	-0.303
Row150	cluster_1	state=mn	-0.412
Row151	cluster_1	state=mo	0.272
Row152	cluster_1	state=ms	0.154
Row153	cluster_1	state=nc	-0.165
Row154	cluster_1	state=ne	-0.44
Row155	cluster_1	state=nj	0.68

Cluster_2 was used as the reference category

Cluster_0 which represent car with maker Ford, model Altima and sedan body

Identity of cluster 0: Customers who value reliability, affordability, comfort, fuel efficiency, and practicality.

We observe that state, color, interior and transmission are most significant variables in cluster 0.

Cluster_1 which represent car with maker Honda, model Camry and SUV body

Identity of cluster 1: Customers who are family-oriented, seeking vehicles that offer ample space and versatility for various activities and lifestyles, also have interest in features that enhance convenience and comfort.

We observe that state, color and interior are most significant variables in cluster 0.

Variables like make, body, trim, selling price, model have no significant impact in distinguishing cluster 1 and cluster 0 from cluster 2.

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_1	17895	0	15636	0	1	1	1	1	1	?	?
cluster_0	7804	0	25727	0	1	1	1	1	1	?	?
cluster_2	7832	0	25699	0	1	1	1	1	1	?	?
Overall	?	?	?	?	?	?	?	?	?	1	1

The overall accuracy of the logistic regression model is very high at 100 and it effectively predicts the cluster labels for the majority of instances. Additionally, the Cohen's Kappa coefficient suggests substantial agreement beyond chance among the predicted and actual cluster labels.

3.2.2.1.2. K-Nearest Neighbours

K=7

Row ID	cluster_1	cluster_0	cluster_2
cluster_1	11970	3029	2896
cluster_0	981	6820	3
cluster_2	927	2	6903

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_1	11970	1908	13728	5925	0.669	0.863	0.669	0.878	0.753	?	?
cluster_0	6820	3031	22696	984	0.874	0.692	0.874	0.882	0.773	?	?
cluster_2	6903	2899	22800	929	0.881	0.704	0.881	0.887	0.783	?	?
Overall	?	?	?	?	?	?	?	?	?	0.766	0.636

K=9

Row ID	cluster_1	cluster_0	cluster_2
cluster_1	11815	3111	2969
cluster_0	791	7010	3
cluster_2	734	2	7096

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_1	11815	1525	14111	6080	0.66	0.886	0.66	0.902	0.757	?	?
cluster_0	7010	3113	22614	794	0.898	0.692	0.898	0.879	0.782	?	?
cluster_2	7096	2972	22727	736	0.906	0.705	0.906	0.884	0.793	?	?
Overall	?	?	?	?	?	?	?	?	?	0.773	0.649

K=19

Row ID	cluster_1	cluster_0	cluster_2
cluster_1	11347	3337	3211
cluster_0	274	7528	2
cluster_2	303	3	7526

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_1	11347	577	15059	6548	0.634	0.952	0.634	0.963	0.761	?	?
cluster_0	7528	3340	22387	276	0.965	0.693	0.965	0.87	0.806	?	?
cluster_2	7526	3213	22486	306	0.961	0.701	0.961	0.875	0.811	?	?
Overall	?	?	?	?	?	?	?	?	?	0.787	0.678

Similarly, we have applied k nearest neighbour for K=11, 13,15,17 and observed that –

In KNN, the number of neighbours to be considered are from k=7 to 19. From the images, it is seen that as the number of k increases the accuracy also increases. For k=19, as the accuracy is the highest from all the other k's, this cluster will be considered.

The overall accuracy of the KNN model is moderate showing mixed performance across different clusters. Cohen's Kappa coefficient also suggests moderate agreement beyond chance among the predicted and actual cluster labels.

3.2.3.1. Model Performance Evaluation of Random Forest

Row ID	cluster_1	cluster_0	cluster_2
cluster_1	17792	0	0
cluster_0	0	7823	0
cluster_2	0	0	7916

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
cluster_1	17792	0	15739	0	1	1	1	1	1	?	?
cluster_0	7823	0	25708	0	1	1	1	1	1	?	?
cluster_2	7916	0	25615	0	1	1	1	1	1	?	?
Overall	?	?	?	?	?	?	?	?	?	1	1

3.3.2. List of Non-Relevant or Unimportant Variables

In the analysis, we see that these were the non-important variables that did not contribute in the supervised learning algorithm which are: -

Car_id, odometer, vin, saledate, selling price , condition and trim.

3.4 List of Relevant or Important Variables

In the analysis, we see that these were the important variables that contributed in the supervised learning algorithm which are: -

Transmission, color, interior, state, make, body and model.

4. Results and Observations

4.1. Comparing Supervised Learning models: Cross Validation using Decision Tree VS Cross Validation using Logistic Regression, KNN

Cross validation using Decision tree

File Hilite

Clusters \ ...	cluster_1	cluster_0	cluster_2
cluster_1	17895	0	0
cluster_0	0	7804	0
cluster_2	0	0	7832

Correct classified: 33,531 Wrong classified: 0

Accuracy: 100% Error: 0%

Cohen's kappa (κ): 1%

Cross validation using Logistic Regression

Clusters \ ...	cluster_1	cluster_0	cluster_2
cluster_1	11930	0	0
cluster_0	0	5202	0
cluster_2	0	0	5222

Correct classified: 22,354 Wrong classified: 0

Accuracy: 100% Error: 0%

Cohen's kappa (κ): 1%

Cross validation using KNN

K=19

Clusters \ ...	cluster_1	cluster_0	cluster_2
cluster_1	11347	3337	3211
cluster_0	274	7528	2
cluster_2	303	3	7526

Correct classified: 26,401 Wrong classified: 7,130

Accuracy: 78.736% Error: 21.264%

Cohen's kappa (κ): 0.678%

Cross Validation using Random Forest

Clusters \ ...	cluster_1	cluster_0	cluster_2
cluster_1	17792	0	0
cluster_0	0	7823	0
cluster_2	0	0	7916

Correct classified: 33,531 Wrong classified: 0

Accuracy: 100% Error: 0%

Cohen's kappa (κ): 1%

5. Managerial Insights

5.1. Appropriate Model

Metrics	Decision Tree	Logistic Regression	KNN	Random Forest
Accuracy (in %)	100%	100%	78.74%	100.00%

The decision tree and logistic regression has the highest accuracy (100%). KNN has significantly lower accuracy of 78.74%.

For this dataset, ensemble learning methods like Random Forest along with Decision Trees with pruning, seem to be the most effective models in terms of accuracy and robustness. Logistic Regression also performs well and provides interpretable results which can be advantageous in certain scenarios. However, KNN appears to be less suitable due to it less accuracy.

Managerial insights according to the appropriate model (Random Forests)

Managerial insights according to the Random Forests model for car prices:

Market Segmentation: Similar to the Decision Tree model, Random Forest can identify customer segments based on car preferences. It can discern nuances in preferences, such as body style, manufacturer, and additional features, leading to targeted marketing efforts.

Dynamic Pricing Strategy: Random Forest's ensemble approach can provide more robust price range estimations compared to a single decision tree. This allows for a dynamic pricing strategy that adapts to changing market conditions and customer preferences. For example, it can account for seasonal trends or fluctuations in demand for specific car models.

Enhanced Inventory Management: By predicting demand with higher accuracy, Random Forest assists in optimizing inventory levels. It enables dealerships to stock the right mix of cars, reducing carrying costs and minimizing the risk of stockouts. This ensures that the dealership meets customer demands efficiently.

Customized Sales Approach: Random Forest's ability to capture complex relationships between car attributes and prices allows for a more personalized sales approach. Sales teams can leverage insights from the model to tailor their pitches to individual customer preferences, enhancing the overall customer experience and increasing sales conversion rates

5.2. Relevant or Important Variables or Features

The relevant or important variables that are used in the decision tree supervised learning algorithm are: -

State

Color

Interior

Transmission

Selling Price

Make

Model