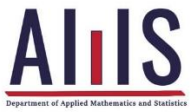
REPORT OF TEAM MOTORCYCLE ANALYSIS

GROUP: AMS-A (1)

Topic : Motorcycle Price Prediction

Institute of Technology of Cambodia

Department of Applied Mathematics and Statistics (AMS)



Topic : Motorcycle Price Prediction

Lecturer : Mr. CHAN Sophal

Members

Name of Student:	ID :
1.Chhon Chaina	e20200934
2.Koythol Amrint	e20201757
3.Kheng Dalish	e20200909
4.Hon Ratana	e20201053
5.Hok Ratanak	e20201106

Contents

1. Introduction

Motorcycles are the convenient mode of transportation which are commonly used in worldwide. Nowadays, there's a massive range of motorbikes available, to suit a wide variety of needs. The popularity of motorcycles has increased significantly and become the key transportation method for many countries, as they allow riding ride through small alleys and traffic congested roads, as well as sporting cheaper prices and better fuel economy than cars. The impact to motorcycle market is mainly linked to the national income or GDP per capita which shows the relative wealth of people in each country. Cambodia is a developing country, whose economic is rapidly increasing and also its population. In order to fulfil people's need of transporting, Cambodia has imported large amount of motorcycles every year. Nowadays, motor market is getting bigger and bigger and as we can see motorcycle on sales in some websites both new and reuse. However, Cambodian people still concern about the price of motorcycle.

The project is aimed to predict the motorcycle market of new and reuse whether it is increase or decrease in the next few years.

II. Exploration Data Analysis

1.1 Data Description

The data was scrapped from website khmer24 which contains information about motors sold in that website. There are 38751 motors on sales and the data shows the motor condition, its location, the prices and also the time sold. Moreover, it describes about motor models and its categories.

- Condition: motor condition whether it is new or reuse
- Price: motor's prices
- Location: where the selling motor is
- Time: the datetime when the motor was put on sales
- Model: motor models
- Category: type of motor
- Year: of as manunfacture of motorcycle

1.2 Data cleaning

2.2.1 Missing values

Missing data are those values that are not present but would have significance if they were. In the real world, the majority of databases have missing data. Data fields with missing data must first be transformed before they can be used for analysis and modeling. To check the missing values of the dataset, we have some method. Within checking the missing values, we can find only 0.09 percent among 38751.

2.2.2 Duplicated

Duplicate data describes situations in which a database or data set contains several instances of the same or extremely similar data. Out of 38751 rows, we were unable to identify any duplicates in the rows.

2.2.3 Outlier

An outlier is a data point in statistics that significantly differs from other observations. An outlier may be caused by measurement variability, a sign of unique data, or an experimental error; the latter is occasionally eliminated from the data set. While an outlier may signal an intriguing possibility, it can also seriously impair statistical analyses.

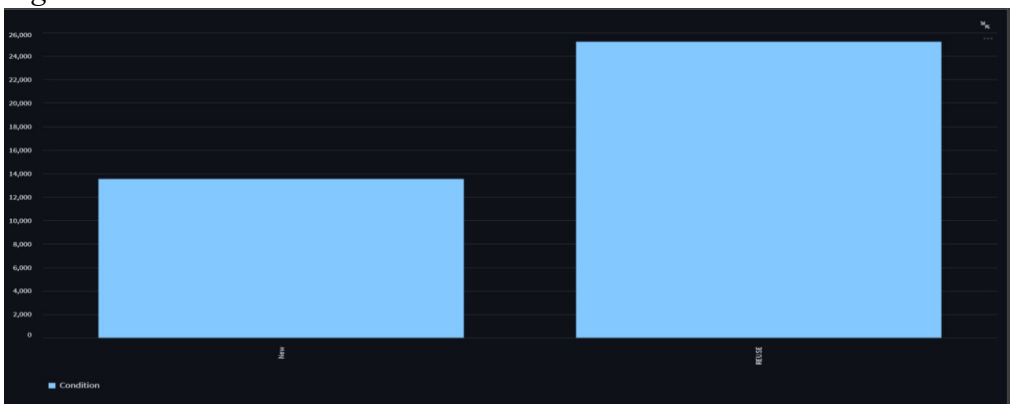
2.3 Data Visualization

For data visualization, we plot histogram, line plot and scatter plot for predicting the number of motor in each location and price depending on models and years.

2.3.1 Condition

Condition refers to situation of motor whether it is new or reuse.

Figure2.3.1

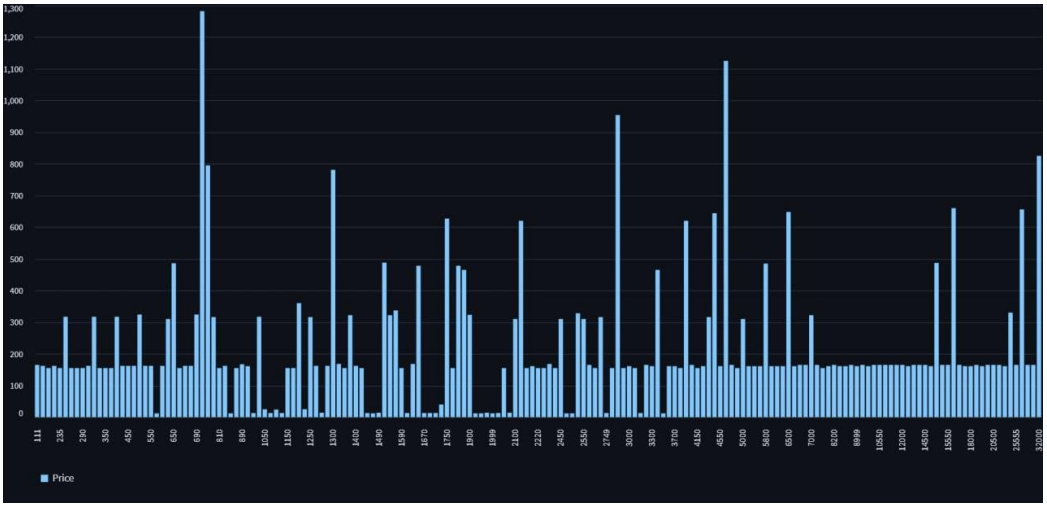


New	Reuse
13,530	25,221

As we can see in figure 2.3.1 the number of reused motors is larger that the number of new motor.

2.3.2 Price

Price of each model depends on its models, conditions and years.

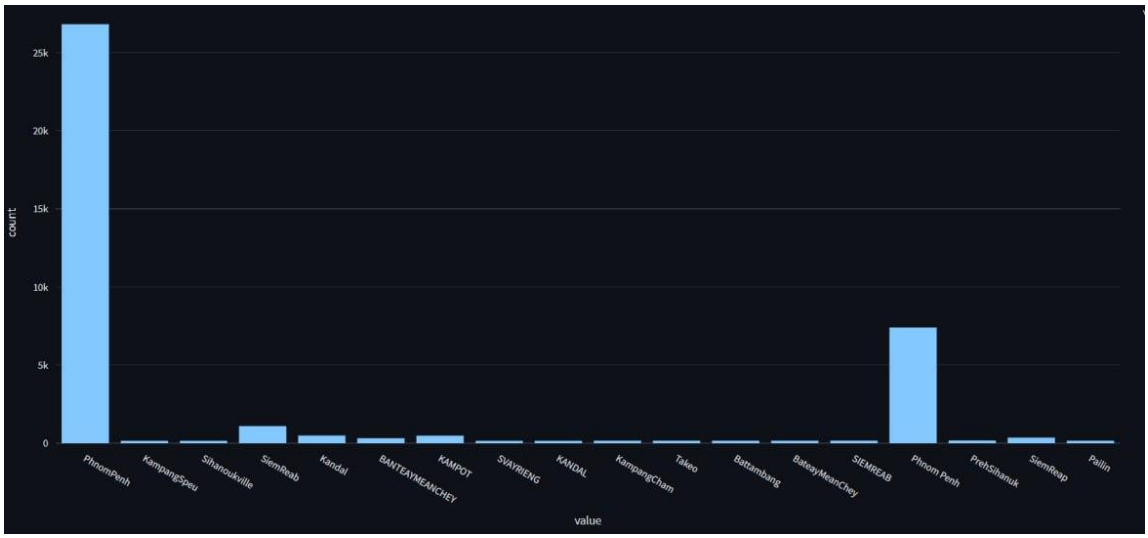


This histogram shows the number of prices of each amount of motor. As we can see, the majority of sales have price $750 with 1282 motors.

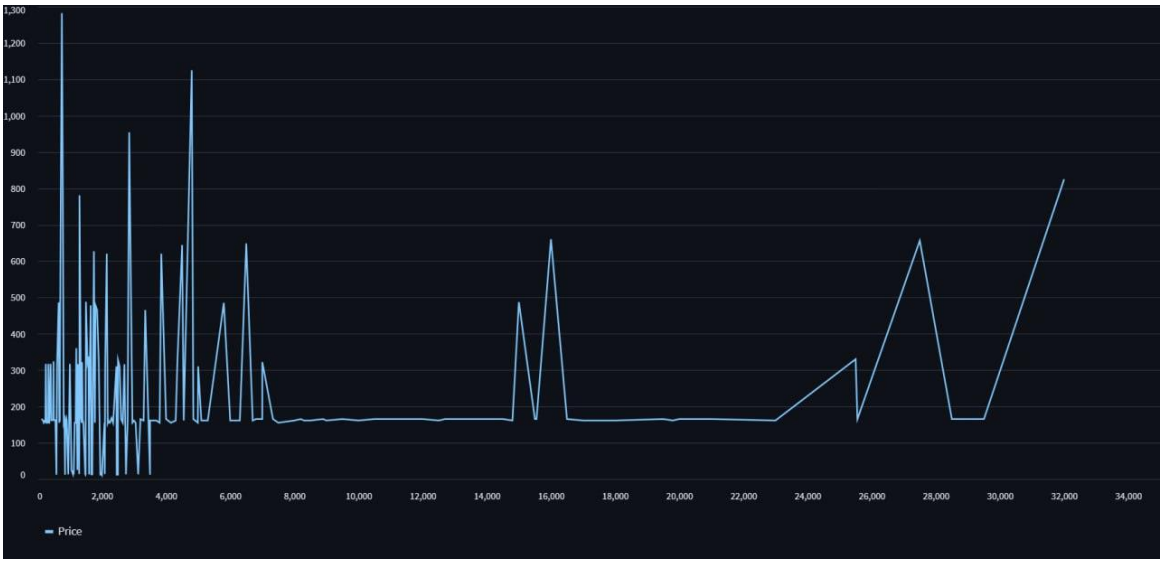
Price	750	4,800	2,850	….	….	1,980	1,485	1,950
Total	1282	1,125	954	….	….	12	12	12

2.2.3 Location

There are different location where motor are sold such as

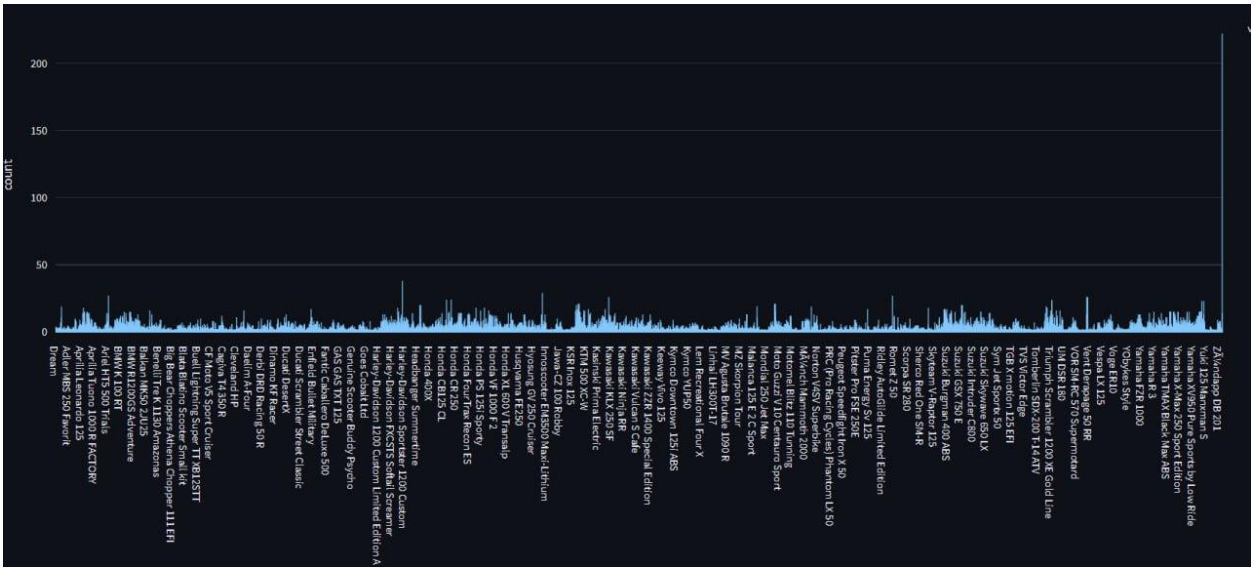


The graph shows about the amount of sales in some provinces; and the highest number is in Phnom Penh.



The line plot illustrates the rate of increasing and decreasing of the price.

2.3.4 Models



The histogram show about motor models

III. Feature engineering

Feature engineering is the process of selecting, transforming, and creating features from raw data to improve the performance of machine learning models. It involves extracting relevant information from the data and representing it in a format that can be effectively utilized by the model. In this case, we divided the data into 2 groups such as numeric features and categorical features. The numerical feature are price and year.

Moreover, the categorical feature has 5 such as Condition, location, time, model and category.

- We find the important feature in our dataset by plotting hit map to the correlation of each feature.
- Identify additional Features they may influence Motorcycle Price such as Model condition, year or geographical location of selling.
- Extract and create relevant features from the available data, ensuring they are in a suitable format for analysis

IV. Models

These classifier provide different approches to solve classification problems and offer flexibility in modeling different types of data. Here's the imported classifiers: Logistic Regression, K-Nearest Neighbors, Gaussian Naive Byes, Decision Tree and Random Forest.

The imported metrics from scikit-learn for evaluating classifier performance : Model, Accuracy, Precision, Recall, F1 Score and F2 Score.

1. Logistic Regression

Model	Accuracy	Precision	Recall	F1 Score	F2 Score
Logistic Regression	0.14153	0.072552	0.14153	0.077486	0.098835

The Logistic Regression model evaluated on the given task has a very low accuracy around 14.15%, indicating that it is not effectively classifying the instances. The precision score of 7.25% suggests that the model is not accurately identifying positive instances, while the recall score of 14.15% indicates that it is not effectively capturing all positive instances in the data. The F1 score and the F2 score suggest that the model's performance on both precision and recall is poor. Overall, these results indicate that the Logistic Regression model has poor performance on this task and may require further exploration or adjustment to improve its effectiveness.

2. K-Nearest Neighbors

Model	Accuracy	Precision	Recall	F1 Score	F2 Score
K-Nearest Neighbours	0.295059	0.308667	0.295059	0.282782	0.286633

These results indicate that the KNN model has poor performance on this task and may require further exploration. The model's accuracy indicates that it classified around 29.51% of instances correctly. The precision score suggests that approximately 30.87% of the instances predicted as positive were true positives. The recall score indicates that around 29.51% of the actual positive instances were correctly identified. The F1 score provides an overall measure of the model's

performance, considering both precision and recall. The F2 score, which places more weight on recall, indicates the model's ability to minimize false negatives.

Comparing the two models, the K-Nearest Neighbors model demonstrates better performance than the Logistic Regression model in terms of accuracy, precision, recall, F1 score, and F2 score. However, the overall performance of both models seems relatively low, and further analysis and improvement might be required to enhance their predictive capabilities.

3. Naive Byes

Model	Accuracy	Precision	Recall	F1 Score	F2 Score
Naive Bayes	0.514772	0.504411	0.514772	0.461412	0.479312

The Naive Bayes model has better performance than the Logistic Regression and KNN models on the given task, with a higher accuracy, precision, recall, and F2 score. However, its F1 score suggests that there is still room for improvement. Overall, the Naive Bayes model appears to be a more effective approach for this task than the other models evaluated.

4. Decision Tree

Model	Accuracy	Precision	Recall	F1 Score	F2 Score
Decision Tree	0.987356	0.988851	0.987356	0.987488	0.987273

The Decision Tree model performs exceptionally well, with high accuracy, precision, recall, F1 score, and F2 score. It indicates that the model has the ability to effectively classify instances and maintain a good balance between precision and recall. The high performance of the Decision Tree model makes it a strong candidate for the given task or problem.

Based on these results, it appears that the Decision Tree model is the most effective approach for this task, with the highest scores across all metrics. The Naive Bayes model also shows promise, but may require further exploration or adjustment to improve its effectiveness. The Logistic Regression and K-Nearest Neighbors models have poor performance and may not be suitable for this particular task.

5. Random Forest

Model	Accuracy	Precision	Recall	F1 Score	F2 Score
Random Forest	0.962844	0.963599	0.962844	0.961732	0.962127

The results suggest that the Random Forest model performs exceptionally well, with high accuracy, precision, recall, F1 score, and F2 score. It indicates that the model has the ability to effectively classify instances and maintain a good balance between precision and recall. The high performance of the Random Forest model makes it a strong candidate for the given task or problem.

The Random Forest and the Decision Tree models have the best performance, with high scores across all metrics. The Naive Bayes model has moderate performance and may be worth further exploration, while the K-Nearest Neighbors and Logistic Regression models have poor performance and may not be suitable for this particular task.

Overall: The Decision Tree model is recommended as the better model among the five options provided.

Model	Accuracy	Precision	Recall	F1 Score	F2 Score
Decision Tree	0.987356	0.988851	0.987356	0.987488	0.987273
Random Forest	0.962844	0.963599	0.962844	0.961732	0.962127
Naive Bayes	0.514772	0.504411	0.514772	0.461412	0.479312
K-Nearest Neighbors (KNN)	0.295059	0.308667	0.295059	0.282782	0.286633
Logistic Regression	0.141530	0.072552	0.141530	0.077486	0.098835

Based on the provided data, it appears to be a classification model evaluation table that includes accuracy, precision, recall, F1 score, and F2 score for different machine learning algorithms. Here's a breakdown of the metrics:

- Accuracy: Represents the overall correctness of the model's predictions.

- Precision: Measures the proportion of correctly predicted positive instances out of the total instances predicted as positive. It indicates the model's ability to avoid false positives.

- Recall: Measures the proportion of correctly predicted positive instances out of the total actual positive instances. It indicates the model's ability to find all positive instances (avoid false negatives).

- F1 Score: It is the harmonic mean of precision and recall, providing a balanced measure of both metrics. It considers both false positives and false negatives.

- F2 Score: Similar to the F1 score, but places more emphasis on recall. It gives more weight to false negatives, which can be useful in certain scenarios.

Looking at the metrics for each model :

- Decision Tree: The model shows high accuracy, precision, recall, F1 score, and F2 score, indicating excellent performance overall.

- Random Forest: The model also performs well, with slightly lower metrics compared to the Decision Tree model but still providing a high level of accuracy and precision.

- Naive Bayes: This model shows significantly lower accuracy, precision, recall, F1 score, and F2 score compared to the other models, suggesting poorer performance.

- K-Nearest Neighbors: The model performs relatively poorly across all metrics, with the lowest accuracy, precision, recall, F1 score, and F2 score among the listed models.

- Logistic Regression: This model performs even worse, with very low accuracy, precision, recall, F1 score, and F2 score, indicating inadequate performance.

Based on these metrics, the Decision Tree and Random Forest models appear to be the most effective among the listed algorithms, while Naive Bayes, K-Nearest Neighbors, and Logistic Regression perform poorly. However, it's important to consider the specific problem, dataset, and requirements when selecting the most suitable model for a given task.

Data Processing

The purpose is to create a data preprocessing pipeline that applies standardization to the input features. The code provided includes some utility functions for evaluating and cross-validating machine learning models using scikit-learn. The function is cross_val(model, X, y), print_evaluate(true, predicted) and evaluate(true, predicted).

The DataFrame structure with the evaluation:

- Model: the names of the models.
- MAE: the mean absolute error (MAE) values.
- MSE: the mean squared error (MSE) values.
- RMSE: the root mean squared error (RMSE) values.
- R2 Square: the R-squared (R2) score values.
- Cross Validation: the cross-validation scores.

1. **Linear Regression**

```
Test set evaluation:
_____________________________________
MAE: 40.31299482323597
MSE: 6399.440461093907
RMSE: 79.99650280539711
R2 Square: 0.9998866997302559
_____________________________________
Train set evaluation:
_____________________________________
MAE: 40.4213385274966
MSE: 5934.974190940419
RMSE: 77.03878368030234
R2 Square: 0.9998969565451178
_____________________________________
```

The results indicate that the model performs well on both the test and training sets. The MAE measures the average absolute difference between the predicted and actual values, with lower values indicating better performance. The MSE measures the average squared difference between the predicted and actual values, and RMSE is the square root of MSE. Again, lower values indicate better performance. The R2 Square score measures the proportion of the variance in the target variable that is predictable from the input features. A value close to 1 indicates a good fit.

The Linear Regression model shows excellent performance, with very low errors and a high R2 Square score close to 1. It captures the relationships between the input features and the target variable very well.

2. **Random Forest Regressor**

```
Test set evaluation:
_____________________________________
MAE: 1.250853954328478
MSE: 196.70022771603686
RMSE: 14.024985836571704
R2 Square: 0.999996517478521
_____________________________________
Train set evaluation:
_____________________________________
MAE: 0.4452880000000011
MSE: 31.891635850193524
RMSE: 5.64726799879318
R2 Square: 0.9999994462950917
_____________________________________
```

The Random Forest model shows excellent performance on both the test and training sets. The high R2 Square values for both train and test sets indicate that the model fits the data well and is able to explain a large proportion of the variance in the target variable. The low values for MAE, MSE, and RMSE indicate that the model has low prediction errors, which suggests that it is making accurate predictions.

These evaluation results suggest that the model is effectively capturing the underlying patterns in the dataset and making accurate predictions.

3.**Logistic Regression**

```
Test set evaluation:
_____________________________________
MAE: 535.0412849954845
MSE: 852291.3065410915
RMSE: 923.1962448694707
R2 Square: 0.9849104252912914
_____________________________________
Train set evaluation:
_____________________________________
MAE: 534.9581290322581
MSE: 861104.7305806451
RMSE: 927.9572892006643
R2 Square: 0.9850494368467712
_____________________________________
```

In logistic regression, the evaluation metrics are slightly different from those in linear regression because it is a classification model. The MAE, MSE, and RMSE values represent the average absolute error, average squared error, and root mean squared error. However, these metrics may not be the most appropriate for evaluating the performance of a logistic regression model.

The R2 Square score can still be useful, a value close to 1 indicates a good fit but it is important to note that R2 Square is not the primary metric used to evaluate classification models.

4.Decision Tree Regressor

```
Test set evaluation:
_____________________________________
MAE: 1.1094052380338022
MSE: 2732.3409882595793
RMSE: 52.2717991679986
R2 Square: 0.9999516246814257
_____________________________________
Train set evaluation:
_____________________________________
MAE: 0.0
MSE: 0.0
RMSE: 0.0
R2 Square: 1.0
_____________________________________
```

The Decision Tree model shows excellent performance on both the test and training sets.

+ Test set evaluation: The Decision Tree model performs well on the test set, with low MAE, MSE, and RMSE values. The R2 Square score is close to 1, indicating that the model explains a significant amount of the variance in the target variable.

+ Train set evaluation: On the training set, the model achieves perfect scores of 0 for MAE, MSE, and RMSE, and an R2 Square score of 1. This suggests that the model perfectly fits the training data, capturing all the patterns and achieving zero error.

These evaluation results suggest that the Decision Tree model has captured the underlying patterns in the data effectively. However, the perfect performance on the training set and relatively low errors on the test set could indicate potential overfitting.

3.Polynomial Regression

```
Test set evaluation:
_____________________________________
MAE: 1.090419728306753 2e-11
MSE: 3.011243713831587e-22
RMSE: 1.735293552639318 3e-11
R2 Square: 1.0
_____________________________________
=====================================
Train set evaluation:
_____________________________________
MAE: 1.09754446677988e-11
MSE: 3.1017197681314 62e-22
RMSE: 1.761169999781810 5e-11
R2 Square: 1.0
_____________________________________
```

The Polynomial Regression model appears to be performing extremely well on both the train and test sets:

+ Test set evaluation: The Polynomial model performs extremely well on the test set. The MAE, MSE, and RMSE values are extremely low, close to zero, indicating that the model's predictions almost perfectly match the actual values. The R2 Square score of 1.0 indicates that the model can explain all the variance in the target variable, resulting in a perfect fit.

+ Train set evaluation: On the training set, the model achieves the same perfect scores of 0 for MAE, MSE, and RMSE, and an R2 Square score of 1.0, indicating a perfect fit to the training data.

These evaluation results suggest that the Polynomial model is able to perfectly capture the underlying patterns in the data, resulting in accurate predictions. However, it is important to consider the possibility of overfitting, especially when the model shows such perfect performance on both the training and test sets.

3. **Support vector machine**

```
Test set evaluation:

_____________________________________
MAE: 3637.5416556173664
MSE: 48939444.89520569
RMSE: 6995.67329820409
R2 Square: 0.13354107418279537

_____________________________________
=====================================
Train set evaluation:

_____________________________________
MAE: 3719.9337434774566
MSE: 50091975.55659756
RMSE: 7077.568477704583
R2 Square: 0.1302994659849085

_____________________________________
```

The SVM model appears to be performing moderately well on both the train and test sets:

+ Test set evaluation: The SVM model's performance on the test set shows relatively high errors. The MAE, MSE, and RMSE values are relatively large, indicating significant differences

between the predicted and actual values. The R2 Square score of 0.133 suggests that the model explains only a small portion (13.35%) of the variance in the target variable, indicating a weak fit to the test data.

+ Train set evaluation: On the training set, the model shows similar performance, with high errors and a low R2 Square score of 0.130. This suggests that the model is not capturing the underlying patterns in the data effectively and is not performing well even on the training data.

These evaluation results suggest that the SVM model may not be appropriate for the given dataset or problem. The model's inability to explain the variance in the target variable and its high errors indicate that it is not accurately capturing the relationships between the input features and the target variable.

V. Model Evolution

Model	Accuracy	Precision	Recall	F1 Score	F2 Score
Decision Tree	0.987356	0.988851	0.987356	0.987488	0.987273
Random Forest	0.962844	0.963599	0.962844	0.961732	0.962127
Naive Bayes	0.514772	0.504411	0.514772	0.461412	0.479312
K-Nearest Neighbors (KNN)	0.295059	0.308667	0.295059	0.282782	0.286633
Logistic Regression	0.141530	0.072552	0.141530	0.077486	0.098835

These metrics provide information about the performance of each model. Higher values for Accuracy, Precision, Recall, F1 Score, and F2 Score generally indicate better model performance. Based on these metrics, the Decision Tree and Random Forest models seem to perform well, while the Naive Bayes, K-Nearest Neighbors, and Logistic Regression models have lower performance scores.

	Model	MAE	MSE	RMSE	R2 Square	Cross Validation
0	Linear Regression	4.031299e+01	6.399440e+03	7.999650e+01	0.999887	0.999897
1	Random Forest Regressor	1.250854e+00	1.967002e+02	1.402499e+01	0.999997	0.999897
2	Logistic Regression	5.350413e+02	8.522913e+05	9.231962e+02	0.984910	0.999897
3	Decision Tree Regressor	1.109405e+00	2.732341e+03	5.227180e+01	0.999952	0.987226
4	Polynomail Regression	1.090420e-11	3.011244e-22	1.735294e-11	1.000000	0.987226
5	Support vector machine(Rergession)	3.637542e+03	4.893944e+07	6.995673e+03	0.133541	0.987226

These metrics provide information about the performance of each regression model. Lower values for MAE, MSE, and RMSE indicate better model performance, while higher values for R2 Square and Cross Validation generally indicate better performance. Based on these metrics, the Linear Regression and Random Forest Regressor models seem to perform well, while the Logistic Regression, Decision Tree Regressor, and Support Vector Machine (Regression) models have lower performance scores. The Polynomial Regression model achieves perfect scores in MAE, MSE, RMSE, and R2 Square, indicating a potentially overfitting model.

V. Conclusion

Based on the above discussion it can be concluded as follows:

- The evaluation of both classification and regression models, the Decision Tree and Random Forest models excel in classifying motorcycle sales data and predicting sales trends. The Linear Regression and Random Forest Regressor models are reliable for accurately predicting motorcycle sales in the Cambodian market. These models can provide valuable insights for market analysis, sales forecasting, and strategic decision-making in the motorcycle industry in Cambodia.
- In conclusion, analyzing the motorcycle sales market in Cambodia requires considering various factors such as motor models, year of manufacture, location, pricing, and motor types. By understanding customer preferences, adapting to market trends, and implementing effective marketing strategies, businesses can position themselves to capitalize on opportunities and drive sales growth in the dynamic Cambodian motorcycle mark.