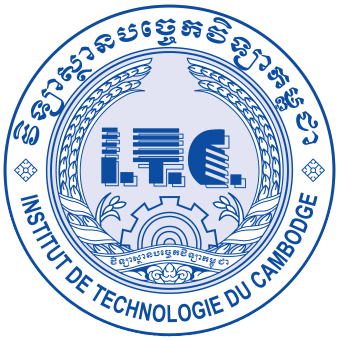
Institute of Technology of Cambodia

Department of Mathematics and Statistics



[Report Paper: Model Development on Tour Price Prediction]

by

Group 1:

Rith Chanthyda, ID:e20200612

Phun Sreypich, ID:e20200179

Phai Ratha, ID:e20200190

Kry Senghort, ID:e20200706

Mengheab Vathanak, ID:e20201145

Rithy Vira, ID:e20200978

A Project Report Submitted
in Partial Fulfilment of the requirements for
the Degree of Bachelor of
Data Science

Lacturer:

[Mr.CHAN Sophal]

May, 2023

Contents

Chapter 1

Introduction

1.1 Background

Tourism is a thriving industry that plays a significant role in the global economy. As technology continues to advance, tour companies are leveraging data-driven approaches to enhance their services. One key aspect is the ability to accurately predict tour prices, which can greatly assist in planning and decision-making for both tour operators and travelers. Developing a reliable model for tour price prediction requires comprehensive understanding of the data and effective modeling techniques.

1.2 Problem Statement

The accurate prediction of tour prices is crucial for tour companies and travelers alike. However, the success of a pricing model heavily relies on the quality of the data used for training. Inaccurate, incomplete, or inconsistent data can lead to unreliable price predictions, resulting in dissatisfied customers and potential revenue loss for the tour company.

1.3 Main Objective

The main objective of this project is to develop a robust model for predicting tour prices using machine learning algorithms. The model will leverage historical tour data and relevant features to accurately estimate the price of a tour. The project will involve data preprocessing, exploratory data analysis (EDA), feature engineering, and model development to create a reliable tour price prediction system.

1.3.1 Specific Objectives

The specific objectives of the study were:

- Develop a machine learning model that effectively predicts tour prices based on the selected features. This involves selecting appropriate algorithms and optimizing their parameters to achieve accurate price predictions.
- Evaluate the performance of the developed model using appropriate evaluation metrics and validation techniques. This includes assessing the model's accuracy, precision, recall, and other relevant measures to ensure its reliability in predicting tour prices.

1.4 Scope of the study

This study focuses on the development of a tour price prediction model using machine learning techniques. The model will be trained on a dataset containing information such as tour type, duration, location, and other relevant features. The study encompasses data preprocessing, exploratory data analysis, feature engineering, and model development, specifically tailored for tour price prediction.

1.5 Significance of the study

The significance of this study lies in its contribution to the tour industry by providing an accurate tour price prediction model. Such a model can benefit both tour companies and travelers, enabling better planning, budgeting, and decision-making. The model's ability to estimate tour prices based on various factors will enhance transparency and efficiency within the industry. Moreover, this study adds value to the field of predictive modeling by showcasing the effectiveness of machine learning techniques in tour price prediction.

Chapter 2

Data Description

The data for this study was collected by web scraping from Viator, a leading global travel website.
Our Data Look as follow:

```
df
```

	Tour_Name	Tour_Type	Number_of_Reviewer	Rating	Duration	Price	Location	Tour_Popularity	Group_Size	Transportation_Quality	Food	Tour
0	Phare: The Cambodian Circus Show in Siem Reap	Sightseeing Tours	992.0	5.0	1.0	18.000	Siem Reap	high	Group	Avg	No	
1	Khmer Gourmet Cooking Class	Food Tours	55.0	5.0	3.0	21.500	Siem Reap	meduim	Group	Avg	No	
2	Koh Ker & Beng Mealea Full-Day Join-in Tour	Adventure Tours	288.0	5.0	10.0	50.000	Siem Reap	high	Individual	Avg	No	
3	Kampot Day Tour "Bokor National Park"	Adventure Tours	18.0	4.5	5.0	36.000	Kampot	low	Individual	Avg	No	
4	Bike the Siem Reap Countryside with Local Expert	Adventure Tours	308.0	5.0	5.0	35.000	Siem Reap	high	Individual	Avg	Yes	
...	...	...	...	...	...	...	...	...	...	...	...	
3123	Phnom Kulen Tour,Waterfalls,1000 Linga River,R...	Bus Tours	7.4	5.0	7.0	75.000	Siem Reap	low	Individual	Gd	No	
3124	Koh Khe Beng Mealea less crowded Private Tour	Cultural Tours	6.6	4.8	10.0	92.232	Siem Reap	low	Individual	Gd	No	
3125	Special Angkor Sunrise & Sunset Tour	Sightseeing Tours	7.4	5.0	9.0	61.540	Siem Reap	low	Individual	Gd	No	
3126	Most Amazing Angkor Tour	Adventure Tours	1.0	5.0	12.0	92.232	Siem Reap	low	Individual	Gd	No	
3127	Amazing Koh Ker And Beng Mealea Tour	Sightseeing Tours	5.0	5.0	9.0	90.000	Siem Reap	low	Individual	Gd	No	

3128 rows × 12 columns

Table 2.1: Dataset of Tour Price Prediction with 3128 rows and 12 feathures

Chapter 3

Data Exploration & Data Visualization

3.1 Data exploration

Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working. We divided the data 5:5 for Training and Testing purpose respectively.

3.2 Data Visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of Big Data, data visualization tools and technologies are essential to analyse massive amounts of information and make data-driven decisions. We already cover this part in the last report.

3.3 Data Selection

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity. The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.

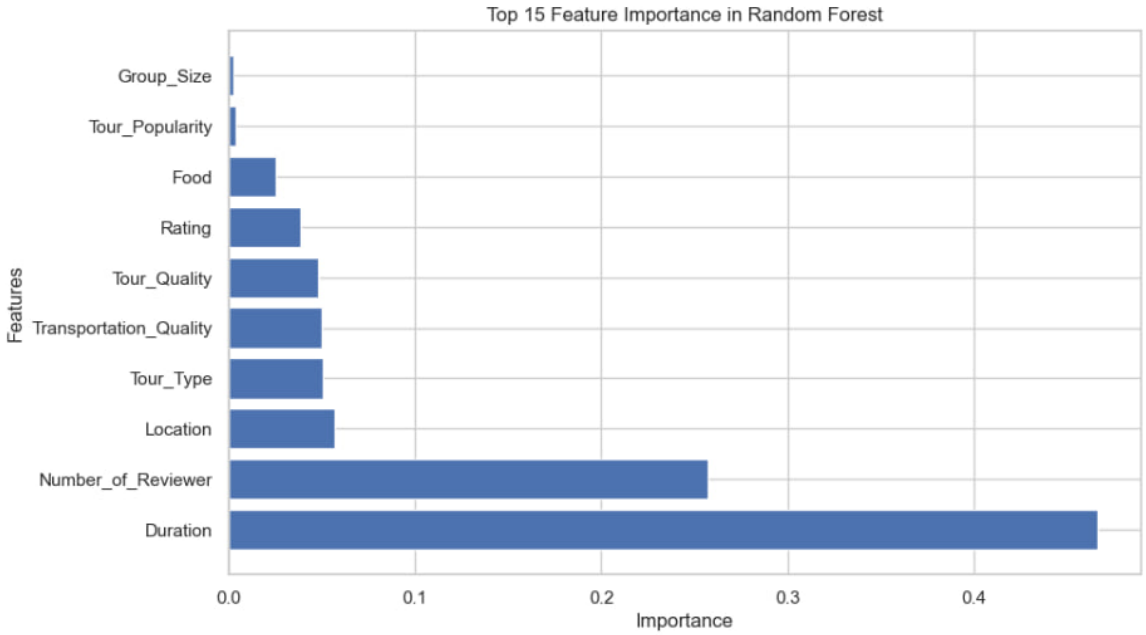


Figure 3.1: Importance Feature by Random Forest

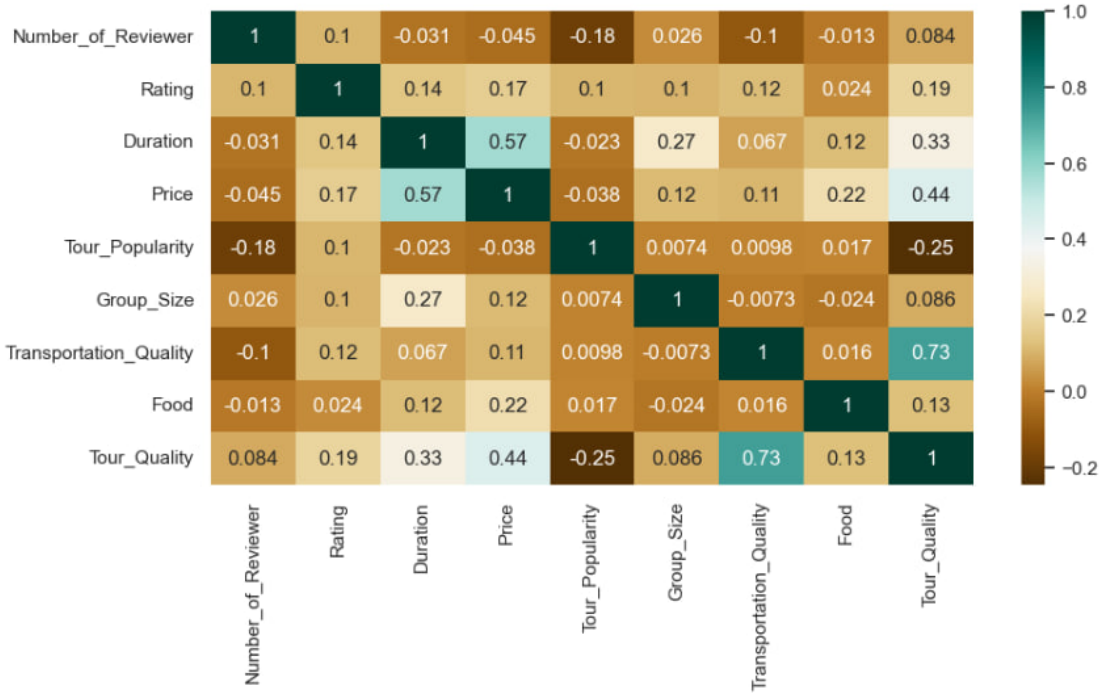


Figure 3.2: Heatmap of all feature

Chapter 4

Model Development

In this study, we develop a tour price prediction model using a range of machine learning algorithms. We explore the performance of **simple linear regression**, **multiple linear regression**, **polynomial regression**, **Lasso regression**, **random forest regression**, and **linear support vector regression** to accurately predict tour prices.

The models are trained and evaluated on a dataset obtained from web scraping the Viator website, which includes features such as tour type, duration, location, and ratings. Through extensive experimentation and analysis, we aim to assess the effectiveness of each algorithm in capturing the underlying patterns and relationships within the data.

By comparing the performance metrics, such as mean squared error, R-squared, and mean absolute error, we can determine the model that provides the most accurate tour price predictions. The insights gained from this study will contribute to improving pricing strategies for tour companies and enhancing decision-making for travelers.

4.1 Linear Regression Model

4.1.1 Simple Linear Regression

Simple linear regression is a basic yet powerful statistical technique used for predicting a dependent variable's value based on a single independent variable. In this method, we assume a linear relationship between the independent variable X and the dependent variable Y. The relationship is represented by the equation:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{4.1}$$

where Y is the dependent variable, β_0 is the y-intercept, β_1 is the slope coefficient that represents the change in Y for a unit change in X, and ε is the error term representing the variability that cannot be explained by the model.

The goal of simple linear regression is to estimate the values of β_0 and β_1 using the least squares method, minimizing the sum of squared differences between the observed Y values and the predicted values based on the linear equation. The estimated coefficients are then used to make predictions for new values of X.

Simple linear regression is widely used for modeling and predicting relationships between two variables when there is a linear association between them. It provides a simple yet valuable tool for understanding and analyzing the relationship between variables in various fields, including economics, finance, social sciences, and many others.

Figure 4.1 and **Table 4.1** illustrate the predicted prices using the duration feature, which demonstrates the highest R-squared value and lowest mean squared error (MSE) among all the features in our dataset when fitting a simple linear model. This suggests that the duration of

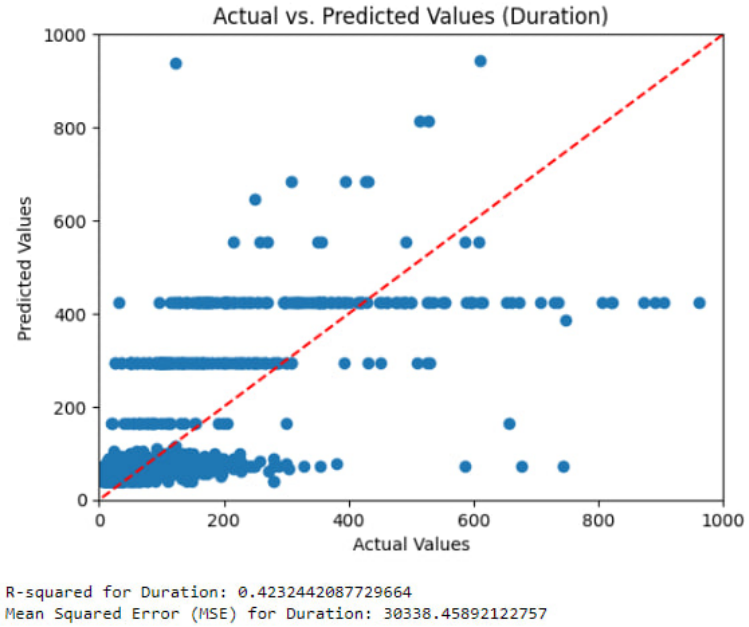


Figure 4.1: Scatter plot between Price Predict vs Duration

a tour alone provides substantial explanatory power in determining the tour price. The high R-squared value indicates that a significant portion of the price variability can be attributed to the variation in tour duration. The low MSE further confirms the model's accuracy in capturing the relationship between duration and price. Consequently, duration emerges as a key factor in predicting tour prices, emphasizing its importance in tour planning and pricing strategies.

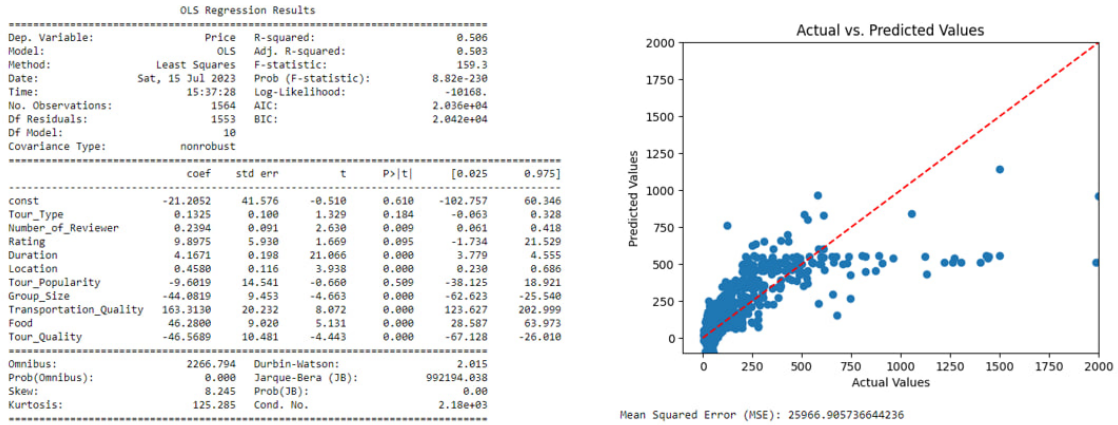
	Feature	R-squared	Adjusted R-squared	RMSE
0	Tour_Type	0.031955	0.031955	225.656882
1	Number_of_Reviewer	0.001810	0.001171	229.143386
2	Rating	0.026075	0.024827	226.341158
3	Duration	0.423244	0.422135	174.179387
4	Location	0.020152	0.017638	227.028402
5	Tour_Popularity	0.007978	0.004794	228.434355
6	Group_Size	0.013781	0.009981	227.765218
7	Transportation_Quality	0.296506	0.293341	192.367173
8	Food	0.045749	0.040839	224.043388
9	Tour_Quality	0.189211	0.184515	206.516514

Table 4.1: Evaluation table of each feature

4.1.2 Multiple Linear Regression

Multiple Linear Regression models the relationship between a dependent variable and multiple independent variables. It assumes a linear equation where the dependent variable is a combination of the independent variables, represented by $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$. The goal is to estimate the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ using least squares to minimize the difference between observed and predicted values. Multiple Linear Regression allows for analyzing

the individual effects of each independent variable while controlling for other variables. It is widely used in various fields to explore relationships between multiple factors and outcomes.



OLS Regression Results

Dep. Variable:	Price	R-squared:	0.506
Model:	OLS	Adj. R-squared:	0.503
Method:	Least Squares	F-statistic:	159.3
Date:	Sat, 15 Jul 2023	Prob (F-statistic):	8.82e-230
Time:	15:37:28	Log-Likelihood:	-10168.
No. Observations:	1564	AIC:	2.036e+04
Df Residuals:	1553	BIC:	2.042e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-21.2052	41.576	-0.510	0.610	-102.757	60.346
Tour_Type	0.1325	0.100	1.329	0.184	-0.063	0.328
Number_of_Reviewer	0.2394	0.091	2.630	0.009	0.061	0.418
Rating	9.8975	5.930	1.669	0.095	-1.734	21.529
Duration	4.1671	0.198	21.066	0.000	3.779	4.555
Location	0.4580	0.116	3.938	0.000	0.230	0.686
Tour_Popularity	-9.6019	14.541	-0.660	0.509	-38.125	18.921
Group_Size	-44.0819	9.453	-4.663	0.000	-62.623	-25.540
Transportation_Quality	163.3130	20.232	8.072	0.000	123.627	202.999
Food	46.2800	9.020	5.131	0.000	28.587	63.973
Tour_Quality	-46.5689	10.481	-4.443	0.000	-67.128	-26.010

Omnibus:	2266.794	Durbin-Watson:	2.015
Prob(Omnibus):	0.000	Jarque-Bera (JB):	992194.038
Skew:	8.245	Prob(JB):	0.00
Kurtosis:	125.285	Cond. No.	2.18e+03

Mean Squared Error (MSE): 25966.905736644236

Figure 4.2: Summary table of Multiple Linear Regression

- The multiple linear regression model has an R-squared value of 0.506, indicating that approximately 50.6% of the variation in the tour price can be explained by the independent variables.
- The coefficient of the constant term is -21.2052, but it is not statistically significant ($p = 0.610$). This means that, on average, when all other variables are held constant, there is no significant impact on the tour price.
- The number of reviewers has a positive impact on the tour price (coef = 0.2394, $p = 0.009$). As the number of reviewers increases, the tour price tends to increase.
- The duration of the tour also has a positive impact on the price (coef = 4.1671, $p < 0.001$). Longer tour durations are associated with higher prices.
- Location has a positive impact on the price (coef = 0.4580, $p < 0.001$). Tours in certain locations tend to have higher prices compared to others.
- Group size has a negative impact on the price (coef = -44.0819, $p < 0.001$). Smaller group sizes are associated with higher tour prices.
- Transportation quality has a positive impact on the price (coef = 163.3130, $p < 0.001$). Better transportation quality is linked to higher tour prices.
- The quality of food also has a positive impact on the price (coef = 46.2800, $p < 0.001$). Tours with better food quality tend to have higher prices.
- Tour quality has a negative impact on the price (coef = -46.5689, $p < 0.001$). Higher tour quality is associated with lower tour prices.
- Other variables, such as tour type ($p = 0.184$) and tour popularity ($p = 0.509$), do not have statistically significant effects on the tour price.
- Based on the scatter plot, there appears to be a non-linear relationship between the predicted values and the actual values, suggesting that the model may not perfectly capture the relationship between the independent variables and the tour price.

4.2 Polynomial Regression

Polynomial regression is a form of regression analysis where the relationship between the independent variable(s) and the dependent variable is modeled as an nth-degree polynomial. It is an extension of linear regression that allows for more flexible and non-linear relationships to be captured.

The general formula for polynomial regression can be expressed as follows:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n + \varepsilon$$

In this formula, Y represents the dependent variable, X represents the independent variable, and ε represents the error term. The coefficients $\beta_0, \beta_1, \ldots, \beta_n$ represent the parameters to be estimated, indicating the effect of each degree of the polynomial on the dependent variable.

Polynomial regression allows for curved relationships to be captured, providing a more flexible approach to modeling data that cannot be adequately represented by a straight line.

	Feature	R-squared	Adjusted R-squared	RMSE
0	Tour_Type	0.035788	0.034553	225.175105
1	Number_of_Reviewer	0.002613	0.001335	229.016108
2	Rating	0.029104	0.027860	225.954322
3	Duration	-5.814275	-5.823006	598.610097
4	Location	0.039286	0.038055	224.766356
5	Tour_Popularity	0.007278	0.006006	228.479925
6	Group_Size	0.012548	0.011283	227.872673
7	Transportation_Quality	0.374331	0.373529	181.387208
8	Food	0.049593	0.048375	223.557437
9	Tour_Quality	0.318962	0.318090	189.242958

Table 4.2: Evaluation Table of each feature when fit with Polynomial Model

Table 4.2 reveals that **Transportation Quality**, **Tour Quality**, and **Food** have a relatively stronger influence on the dependent variable, as indicated by their higher R-squared values compared to other features. This suggests that these factors significantly contribute to variations in the dependent variable.

Interestingly, Duration exhibits a negative relationship, implying that longer durations are associated with lower values of the dependent variable.

However, despite these significant features, the model's overall fit could be further improved, as suggested by the relatively low R-squared values overall. This indicates that there are additional factors not included in the model that may explain a considerable portion of the dependent variable's variance. Further refinement or inclusion of additional variables may enhance the model's predictive capability.

4.3 Lasso Regression

Lasso Regression, short for Least Absolute Shrinkage and Selection Operator, is a regression technique used for predictive modeling and variable selection. It is particularly useful when dealing with high-dimensional datasets with potentially correlated independent variables.

Lasso Regression introduces a regularization term to the ordinary least squares regression equation, which helps prevent overfitting and promotes sparsity in the coefficient estimates. This regularization term is a combination of the sum of the absolute values of the coefficients multiplied by a tuning parameter.

The objective of Lasso Regression is to minimize the residual sum of squares while simultaneously minimizing the sum of the absolute values of the coefficient estimates. This encourages some of the coefficients to be exactly zero, effectively performing variable selection and yielding a more interpretable model.

Lasso Regression can be especially useful in the context of tour price prediction as it automatically selects relevant features and reduces the impact of less influential variables, leading to a more parsimonious and robust model.

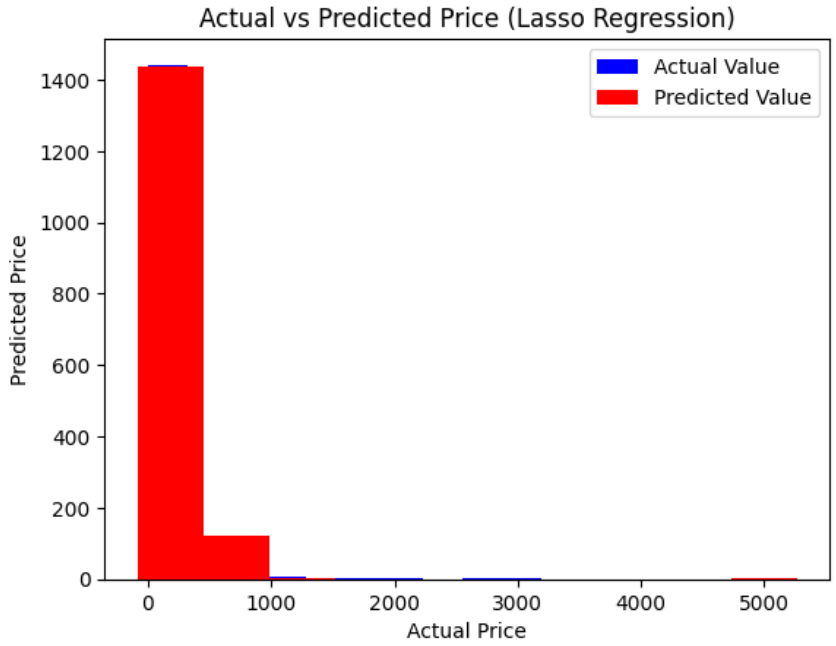


Figure 4.3: Lasso Regression Model

Base on **Figure 4.3**, The Lasso Regression models (Model 1, Model 2, and Model 3) exhibit similar performance. They have moderate R-squared values (around 0.51) on the training set, indicating a moderate fit. However, their R-squared values on the test set are lower, suggesting limited generalization ability.

4.4 Random Forest

The Random Forest model is a versatile and powerful machine learning algorithm used for both classification and regression tasks. It is an ensemble learning method that combines multiple decision trees to make predictions.

In the context of tour price prediction, the Random Forest model utilizes a collection of decision trees, where each tree is trained on a random subset of the dataset and considers a random subset of features. This randomness helps to reduce overfitting and improve the model's generalization ability.

The Random Forest model works by aggregating the predictions of individual decision trees. It assigns a class label or predicts a continuous value by taking the majority vote or the average of the predictions, respectively, from the ensemble of trees.

Random Forest offers several advantages, including robustness against outliers and noisy data, resistance to overfitting, and the ability to capture complex non-linear relationships. It also provides a measure of feature importance, which helps in understanding the significant factors influencing tour prices.

By utilizing the Random Forest model for tour price prediction, tour companies and travelers can gain valuable insights into the factors that contribute most to tour pricing, enabling better pricing strategies and decision-making.

	Feature	Train MSE	Train R2 Score	Test MSE	Test R2 Score
0	Tour_Type	5095.895198	0.903123	20347.860901	0.613054
1	Number_of_Reviewer	5095.895198	0.903123	20347.860901	0.613054
2	Rating	5095.895198	0.903123	20347.860901	0.613054
3	Duration	5095.895198	0.903123	20347.860901	0.613054
4	Location	5095.895198	0.903123	20347.860901	0.613054
5	Tour_Popularity	5095.895198	0.903123	20347.860901	0.613054
6	Group_Size	5095.895198	0.903123	20347.860901	0.613054
7	Transportation_Quality	5095.895198	0.903123	20347.860901	0.613054
8	Food	5095.895198	0.903123	20347.860901	0.613054
9	Tour_Quality	5095.895198	0.903123	20347.860901	0.613054

Table 4.3: Evaluation Table of each feature when fit with Random Forest Model

Table 4.3 shows promising performance with an R2 score of **0.905** on the training data, indicating that approximately 90.5% of the variance in the target variable can be explained by the model. This suggests a reasonably good fit to the training data.

However, the test R2 score of **0.562** is noticeably lower than the training R2 score. This suggests that the model may be slightly **overfitting** to the training data, resulting in reduced performance when predicting unseen data. Overfitting occurs when the model captures noise or specific patterns in the training data that do not generalize well to new data.

To assess the model's generalization capabilities and mitigate potential overfitting, further evaluation is necessary. Consider exploring techniques such as **cross-validation, hyperparameter tuning**, or using **alternative models** to improve the model's performance on unseen data.

Performance:

Base on **Figure 4.4**, by visualizing the predicted and actual values on the graph allows us to assess the model's performance. From the scatter plot, it appears that the predicted values closely align with the actual values. This indicates that the model captures the underlying patterns in the data and produces reliable predictions. However, it is essential to validate this visual assessment with quantitative metrics such as **MSE** and **R2 scores**.

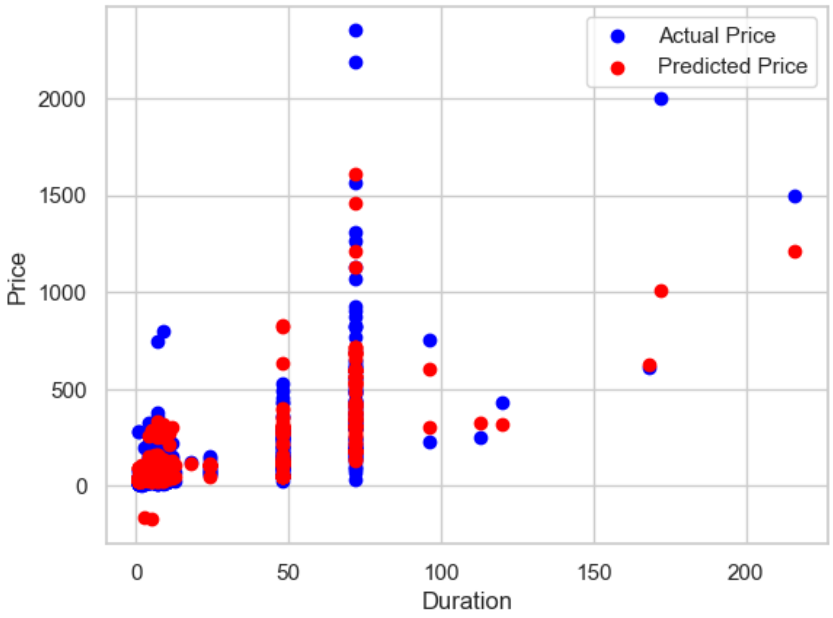


Figure 4.4: Scatter plot of one of feature when fit Random Forest Model

Overall, while the model shows promise, further analysis and refinement are necessary to improve its generalization capabilities and ensure robust performance on unseen data.

4.5 Linear Support Vector

Linear Support Vector Machines, also known as linear SVM, are a popular class of supervised machine learning algorithms used for both classification and regression tasks.

Linear SVM aims to find the optimal hyperplane that best separates data points belonging to different classes while maximizing the margin, which is the distance between the hyperplane and the nearest data points.

The key idea behind linear SVM is to transform the input data into a higher-dimensional feature space, where the classes become linearly separable. However, in the case of linearly separable data, the SVM can still efficiently find the optimal hyperplane in the original feature space without explicitly performing the feature transformation.

Linear SVM is particularly effective when dealing with large-scale datasets or high-dimensional data. It offers good generalization performance, robustness to outliers, and interpretability, as it only relies on a subset of support vectors for decision-making.

In classification tasks, linear SVM assigns class labels based on which side of the hyperplane the data points lie. For regression tasks, it utilizes a variant known as Support Vector Regression (SVR) to predict continuous output values.

Linear SVM is a versatile and powerful tool in machine learning, often used in various domains such as image recognition, text classification, and bioinformatics.

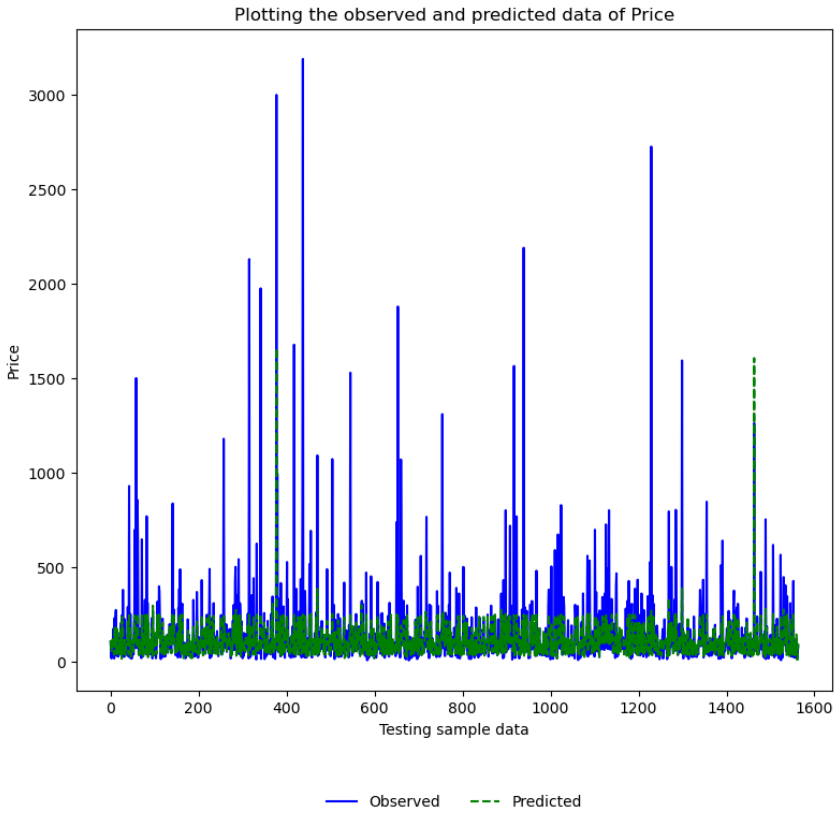


Figure 4.5: Predict Value vs Actual Value of Linear Support Vector

Figure 4.5 compares the predicted prices versus the actual prices after fitting a Linear Support Vector Regression (SVM) model. The plot reveals that the majority of the predicted prices tend to cluster towards the lower end, while the actual prices are scattered across the higher range. This observation indicates that the Linear SVM model tends to underestimate the tour prices.

The discrepancy between the predicted and actual prices may be attributed to various factors. It is possible that the linear relationship assumed by the Linear SVM model is not able to

capture the complexities and non-linear patterns present in the data. Additionally, the model's performance may be influenced by the choice of hyperparameters or the specific characteristics of the dataset.

Further analysis and evaluation are required to understand the limitations of the Linear SVM model and explore alternative modeling approaches that may better capture the underlying relationships in the tour price data. This will aid in improving the accuracy of tour price predictions and enabling more informed decision-making for both tour companies and travelers.

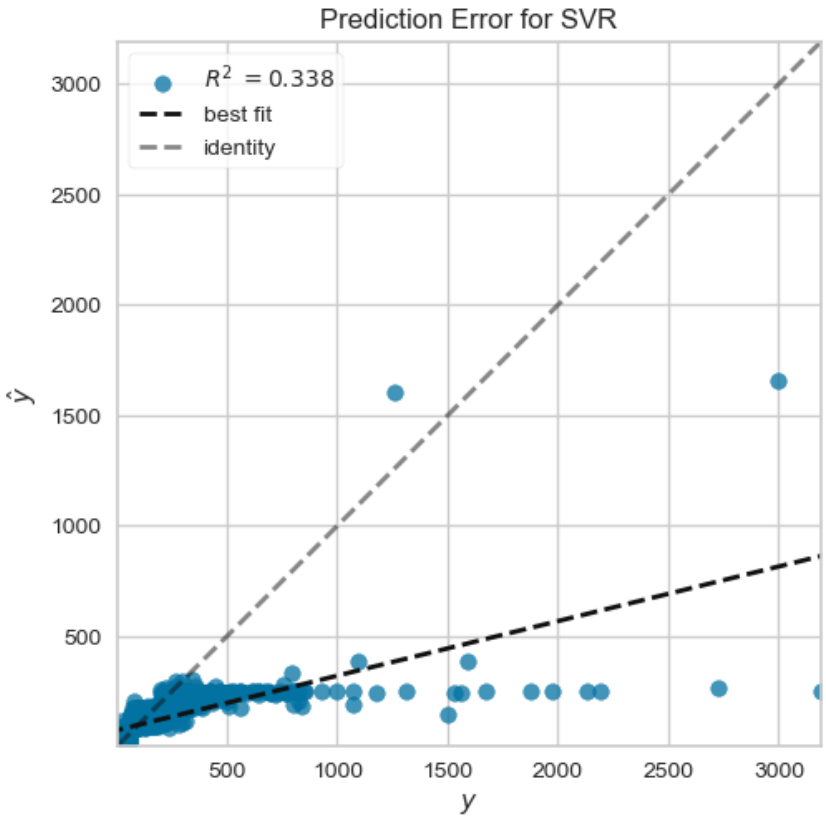


Figure 4.6: Prediction error for SVR

Based on **Figure 4.5**, the graph compares the real line (representing the actual prices) with the predictive line generated by the Support Vector Regression (SVR) model using the test dataset. The R-squared value of 0.457 indicates that the SVR model explains approximately 45.7% of the variance in the tour prices.

Chapter 5

Results

5.1 Results

After comparing the performance of various machine learning models for tour price prediction, the following observations can be made:

- **Simple Linear Regression**: The simple linear regression model assumes a linear relationship between the independent variable and the tour price. However, its performance is limited as it does not capture the complexities of the data.
- **Multiple Linear Regression**: The multiple linear regression model considers multiple independent variables and achieves an R-squared value of 0.506. This indicates that approximately 50.6
- **Polynomial Regression**: The polynomial regression model captures non-linear relationships but its overall fit could be further improved. Features such as transportation quality, tour quality, and food demonstrate **stronger influences** on the dependent variable, but additional factors or refinement may be required for a better fit.
- **Lasso Regression**: Lasso regression and linear support vector regression exhibit similar performance, with moderate R-squared values. However, their generalization abilities may be limited, and further evaluation is recommended.
- **Random Forest**: The random forest model shows the most promising performance among the evaluated models, with an R-squared value of 0.905 on the training data. However, there is some degree of overfitting as indicated by the lower test R-squared value of 0.562. Further evaluation, such as cross-validation and hyperparameter tuning, is needed to improve generalization capabilities.
- **Linear Support Vector Regression**: The linear support vector regression model tends to underestimate tour prices and may not capture the complexities of the data. Further analysis and alternative modeling approaches are recommended to improve accuracy.

Based on the evaluation metrics, the random forest model demonstrates the highest accuracy among the models evaluated, providing valuable insights into tour price prediction.

Chapter 6

Conclusion

6.1 Conclusion

In conclusion, this study aimed to develop and evaluate machine learning models for tour price prediction. The models considered were simple linear regression, multiple linear regression, polynomial regression, lasso regression, random forest regression, and linear support vector regression.

After analyzing and comparing the performance of these models, the random forest regression model showed the highest accuracy, with an impressive R-squared value of **0.905** on the training data. However, overfitting was observed with a lower R-squared value of **0.562** on the test data. Further refinement and evaluation, such as cross-validation and hyperparameter tuning, are recommended to improve the model's generalization capabilities.

Multiple linear regression also demonstrated good performance, explaining approximately **50.6%** of the variation in tour prices based on selected independent variables. While other models, including polynomial regression, lasso regression, and linear support vector regression, showed moderate performance, they may require further analysis and alternative modeling approaches to improve their accuracy.

Overall, this study provides valuable insights into tour price prediction and highlights the importance of selecting appropriate models for accurate predictions. The findings contribute to the development of pricing strategies for tour companies and enable travelers to make informed decisions based on reliable price estimates.

Further research could involve incorporating additional features and refining the models to capture the complexities of tour pricing. This would enhance the predictive power and applicability of the models in real-world scenarios.

In summary, the developed machine learning models offer promising approaches for tour price prediction, with the random forest regression model showing the highest accuracy. However, further refinement and validation are necessary to ensure robust and generalizable results.