Table of Contents

[Table of Figures 3](#_Toc186141195)

[Introduction 4](#_Toc186141196)

[Exploratory Data Analysis 5](#_Toc186141197)

[Missing Values 5](#_Toc186141198)

[Imputation of Missing Values 5](#_Toc186141199)

[Renaming Columns 6](#_Toc186141200)

[Outliers Detection 6](#_Toc186141201)

[Creation of Features 6](#_Toc186141202)

[Encoding 7](#_Toc186141203)

[Scaling 7](#_Toc186141204)

[Feature Engineering 7](#_Toc186141205)

[Correlation 7](#_Toc186141206)

[Feature Selection 7](#_Toc186141207)

[Dimensional Reduction 8](#_Toc186141208)

[PCA 8](#_Toc186141209)

[LDA 9](#_Toc186141210)

[Dimensional Reducer Conclusion 10](#_Toc186141211)

[Statistical Analysis 11](#_Toc186141212)

[Descriptive Statistics 12](#_Toc186141213)

[Shapiro Wilk Test 13](#_Toc186141214)

[Population Samples 14](#_Toc186141215)

[Anova Test 14](#_Toc186141216)

[T-test 15](#_Toc186141217)

[Covariance 16](#_Toc186141218)

[Correlation 17](#_Toc186141219)

[Regression 18](#_Toc186141220)

[Machine Learning Model 19](#_Toc186141221)

[Classifier Models 20](#_Toc186141222)

[30% Split Test 20](#_Toc186141223)

[15% Split Test 21](#_Toc186141224)

[Hyperparameters Tunning and Cross Validation 23](#_Toc186141225)

[Classification Model Conclusion 25](#_Toc186141226)

[Regressor Models 26](#_Toc186141227)

[Hyperparameters Tunning and Cross Validation 27](#_Toc186141228)

[Regression Model Conclusion 27](#_Toc186141229)

[Conclusion 29](#_Toc186141230)

[References 30](#_Toc186141231)

[Github 30](#_Toc186141232)

# Table of Figures

[Figure 1 Missing Values in each column 5](#_Toc186141140)

[Figure 2 Feature Importance for Rider Satisfaction 8](#_Toc186141141)

[Figure 3 Cumulative Variance by Principal Components 9](#_Toc186141142)

[Figure 4 Cumulative Variance by Linear Discriminations 10](#_Toc186141143)

[Figure 5 Contingency Table for Weather Condition and City 11](#_Toc186141144)

[Figure 6 Contingency Table for City and Bike Model 11](#_Toc186141145)

[Figure 7 Contingency Table for Ride Satisfaction and Bike Model 12](#_Toc186141146)

[Figure 8 Shapiro-Wilk Test Results 13](#_Toc186141147)

[Figure 9 Satisfaction Category Samples & Confidence Interval 14](#_Toc186141148)

[Figure 10 Anova Test for Satisfaction Categories 15](#_Toc186141149)

[Figure 11City Samples & Confidence Interval 15](#_Toc186141150)

[Figure 12 T-Test Results for Berlin and New York 16](#_Toc186141151)

[Figure 13 T-Test Results for Berlin and London 16](#_Toc186141152)

[Figure 14 Covariance Matrix 16](#_Toc186141153)

[Figure 15 Correlation Matrix 17](#_Toc186141154)

[Figure 16 Regression Model Results 18](#_Toc186141155)

[Figure 17 Classifier Models Integration 20](#_Toc186141156)

[Figure 18 Classifier Models Result for Original Dataset and 30% Split 20](#_Toc186141157)

[Figure 19 Classifier Model Results for the Scaled Dataset and 30% Split 20](#_Toc186141158)

[Figure 20 Classifier Model Results for PCA Reduced Dataset and 30% Split 21](#_Toc186141159)

[Figure 21 Classifier Model Results for LDA Reduced Dataset and 30% Split 21](#_Toc186141160)

[Figure 22Classifier Model Results for Original Dataset and 15% Split 21](#_Toc186141161)

[Figure 23 Classifier Model Results for the Scaled Dataset and 15% Split 22](#_Toc186141162)

[Figure 24 Classifier Model Results for PCA Reduced Dataset and 15% Split 22](#_Toc186141163)

[Figure 25 Classifier Model Results for LDA Reduced Dataset and 15% Split 22](#_Toc186141164)

[Figure 26 Random Forest Report 23](#_Toc186141165)

[Figure 27 Random Forest Confusion Matrix 23](#_Toc186141166)

[Figure 28 10-Fold Cross-Validation for Random Forest 24](#_Toc186141167)

[Figure 29 K-Nearest Neighbours Report 24](#_Toc186141168)

[Figure 30 K-Nearest Neighbours Confusion Matrix 25](#_Toc186141169)

[Figure 31 10-Fold Cross-Validation Results for KNN Model 25](#_Toc186141170)

[Figure 32 Regression Models 26](#_Toc186141171)

[Figure 33 Regression Model Results for Original Dataset and 30% Split 26](#_Toc186141172)

[Figure 34 Random Forest Results After Hyperparameters Tunning 27](#_Toc186141173)

[Figure 35 10-Folds Cross-Validation Results 27](#_Toc186141174)

# Introduction

Data analysis is a crucial tool for understanding patterns, relationships and trends within a dataset. This report presents an exploratory analysis and predictive modelling performed on a bicycle rental dataset. Through advanced techniques such as missing value imputation, feature engineering, dimensionality reduction and statistical analysis, we seek to understand the factors that impact user satisfaction and associated trip costs. In addition, classification and regression models were evaluated to predict both user satisfaction and rental cost, exploring various pre-processing and model fitting configurations. This paper not only highlights the limitations and challenges of the dataset, but also offers practical recommendations based on the findings.

# Exploratory Data Analysis

The dataset has been found in XSLX version; a Python code has been created to transform the document to csv in order to start working with it.

The dataset has a size of 8 columns and 900 observations. It contains a composition of four categorical variables and four numerical variables.

## Missing Values

The dataset contains a range of at least 1 - 1.7% missing values for each column over the total observations. The following image displays how many missing values each column contains:

A screenshot of a computer

Description automatically generated

Figure Missing Values in each column

## Imputation of Missing Values

A function ‘fill\_missing\_values’ has been created that obtains as parameters the dataset, the column to fill in the missing values, and the columns to be taken as reference. The purpose of this function is to fill in the missing values of the column (target\_column) of the Dataset, using information from other columns (reference\_columns) to find patterns and fill in the missing values.

The function will determine whether the variable is categorical or numeric as it will use the median for numeric and mode for categorical. The goal is to find ‘neighbours’ with similar characteristics to infer the missing value.

To fill in the missing values for the variables in the dataset, those variables that were considered most related or relevant to the target variable were selected as a reference. This selection was based on the idea that these variables could provide useful information to estimate the missing values in a consistent and accurate manner.

* City: Weather Condition, Bike Model, and Distance Covered (km) have been used.
* Weather Condition: City, Rider Satisfaction, and Ride Duration (min) have been used.
* Rider Age: City, Bike Model, and Distance Covered (km) have been used.
* Bike Model: City, Ride Duration (min), Distance Covered (km), Rider Satisfaction have been used.
* Ride Duration (min): Rider Age, Weather Condition, Distance Covered (km), Bike Model have been used.
* Distance Covered (km): Rider Age, Weather Condition, Ride Duration (min), and Bike Model have been used.
* Rider Satisfaction: Ride Duration (min), Weather Condition, Distance Covered (km), and Bike model have been used.
* Bike Rental Cost ($): Ride Duration (min), City, Distance Covered (km), and Bike Model have been used.

After using the function, it has been confirmed that there were no missing values and no duplicate data.

## Renaming Columns

Most of columns have been renamed for better readability and understanding. The following columns have been renamed:

* "Ride Duration (min)": "Duration"
* "Distance Covered (km)": "Distance",
* "Bike Rental Cost ($)": "Cost"}

## Outliers Detection

Different plots have been created in order to detect any outliers; they confirmed that each column does not contain outliers.

## Creation of Features

## Five new variables were created in order to be able to extract more information than the dataset presents. Among them are:

## Age Category: To classify the age of the people.

## Speed (km/h): Indicates the average speed of the trip.

## Speed Category: To classify the speeds.

## Cos\_per\_km: The cost per kilometre travelled.

## Cost\_per\_minute: The cost per minute travelled.

## Encoding

The variables City, Weather Condition, Bike Model, Age Category and Speed\_category were taken, and one hot encoding was applied to them since these variables do not have ordinal values, in this case they are nominal values.

For the target variable (Rider Satisfaction) label encoder was applied since this variable does have nominal values and the variable cannot be separated.

## Scaling

Two types of scalers were used since both have benefits for the analysis.

* MinMax scaler: Since the dataset has no outliers, this scaler is perfect since the range cannot be distorted, and scales can be obtained in a fixed range between 0 and 1. MinMax Scaler was applied to Rider Age, Duration, Distance, Cost, Speed (km/h), Cost\_per\_km, and Cost\_per\_minute as these variables have continuous values that can be in different ranges, which makes scaling useful.
* StandardScaler: This scaler was used because PCA (Principal Analysis Component) will be used since it is ideal for this algorithm which assumes that the data are normally distributed, (Jaadi, Z, 2024). This means that the data will have a mean of 0 and a standard deviation of 1. Therefore, the StandardScaler has been applied to the same variables as in MinMax.

## Feature Engineering

### Correlation

The correlation matrix suggests that the Rider Satisfaction variable does not show a significant relationship with the other variables, which could be because it is influenced by other factors not included in the data or there may be more complex patterns that a simple correlation does not capture.

### Feature Selection

A random forest model was used to obtain the feature importance of the dataset with respect to rider satisfaction. The following image shows the results obtained.

A screenshot of a computer

Description automatically generated

Figure Feature Importance for Rider Satisfaction

The variables that will be used to feed the model in its first phase are: Cost, Cost\_per\_minute, Speed (km/h), Distance, Cost\_per\_km , Duration, Rider Age since they are the most significant in the analysis. A decision tree model was used which resulted in an accuracy of 34.8%.

## Dimensional Reduction

### PCA

Since the dataset contains 26 columns and it was observed that some variables presented redundancy and collinearity, it was decided to apply Principal Component Analysis (PCA) as a dimensionality reduction technique. The objective was to simplify the model by reducing the number of variables without losing a significant amount of information.

To determine the optimal number of principal components, the cumulative proportion of variance explained by each principal component was analysed. The criterion was to retain 99.5% of the total variance to ensure that most of the relevant information from the original dataset was preserved.

As a result, 20 principal components were found to be sufficient to explain 99.5% of the variance. This means that the model succeeded in reducing the number of dimensions from 26 to 20, while retaining virtually all the significant information from the original dataset.

After performing dimensionality reduction using PCA, a Decision Tree Classifier was applied to evaluate the effectiveness of the reduced data. The classifier achieved an accuracy of 0.359, indicating that while the PCA retained most of the variance, additional tuning or feature engineering might still be required to improve the model's performance.

A graph with a line

Description automatically generated

Figure Cumulative Variance by Principal Components

### LDA

The target variable, Rider Satisfaction, was treated as the categorical variable for the LDA transformation. LDA was fitted to the scaled data, and the explained variance ratio for each linear discriminant was calculated. A threshold of 99.5% cumulative variance was used to determine the optimal number of components.

Based on the analysis, 2 linear discriminants were sufficient to retain 99.5% of the variance in the data. This represents a significant reduction in dimensionality from the original dataset, which consisted of 26 features, down to just 2 components.

A graph with a line

Description automatically generated

Figure Cumulative Variance by Linear Discriminations

After dimensionality reduction, a decision tree classifier was applied to the transformed data to evaluate the performance. The accuracy of the model after applying LDA was 0.344, suggesting that while dimensionality reduction was effective, the linear discriminants may not fully capture the complexity required for high prediction accuracy.

### Dimensional Reducer Conclusion

PCA is better suited for clustering and exploratory analysis, while LDA is more appropriate for classification tasks due to its ability to maximize class separability. Both techniques reduced dimensionality effectively, but their applications and impacts on predictive tasks differ significantly.

# Statistical Analysis

Contingency tables were chosen to identify relationships of some variables, for example:

City vs Weather Condition

A screenshot of a computer

Description automatically generated

Figure Contingency Table for Weather Condition and City

Rainy weather (252 rides) is the most frequent condition overall, followed closely by Sunny weather (222). This could suggest that rainy conditions are prevalent in the cities analyzed. Cloudy and Snowy weather occur less frequently but still show significant counts, particularly in cities like Berlin and London.

City vs Bike Model

A screenshot of a computer

Description automatically generated

Figure Contingency Table for City and Bike Model

Usage is evenly distributed across bike models. This balance suggests a diverse set of preferences or needs among riders. In addition, Electrical bike usage is balanced across most cities, with totals ranging from 56 to 61 in Berlin, Chicago, London, and New York. Regarding Mountain bikes, they are more common in cities like London and Berlin, it could be related to terrain (e.g., more parks or hilly areas) or rider preferences. For standard bike popularity in New York may indicate infrastructure or shorter travel distances that suit this model better.

Bike Model vs Rider Satisfaction: Understand how satisfaction levels differ by bike model.

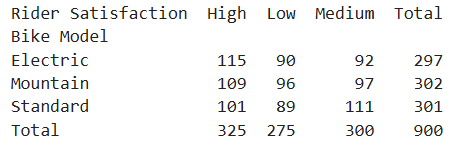


Figure Contingency Table for Ride Satisfaction and Bike Model

High Satisfaction: This suggests a potential advantage in rider comfort or functionality for Electric bikes.

Low Satisfaction: Mountain bikes have the highest number of low satisfaction riders (96), which might point to issues like usability or suitability.

Medium Satisfaction: Standard bikes dominate this category (111), possibly reflecting average performance or user experiences that don't exceed expectations.

## Descriptive Statistics

1. Rider Age

The age distribution shows a broad range from young adults (18) to older adults (60). Most riders fall between 29 and 50 years, as indicated by the IQR and the average rider age (39 years) suggests that mid-aged individuals are the most common users of the service.

Finally, it can be concluded that Rider Age is likely to have an approximately normal distribution, as the symmetry and centrality of the mean with respect to the median are indicative of this.

1. Duration

A standard deviation of 33.89 suggests significant variation in trip durations. Most trips are concentrated between 32 and 92 minutes and a few very short trips (5 minutes) and very long trips (120 minutes) could represent outliers or unique scenarios (e.g., accidental trips, long commutes).

It can be concluded that although the mean and median are close, the high standard deviation and the wide range (5 to 120) suggest that the Duration variable might not be completely normal and have some positive skewness.

1. Distance

The average trip distance aligns closely with the median, indicating a relatively symmetrical distribution of distances. The minimum value (0.5 km) might suggest extremely short or edge-case trips (e.g., test rides) and most trips (IQR) are between 10.87 and 29.12 km, reflecting typical urban or suburban commutes. The Distance variable appears to have an approximately normal distribution due to the closeness of the mean and median, as well as a moderate dispersion. The Cost variable may not follow a perfectly normal distribution, as the mean and median are not identical, and a slight skew may exist.

1. Cost

Strong correlation with Distance and Duration: Higher costs are likely linked to longer trips (distance/duration). Most trips cost between 19.25 and 34.08 and trips with costs as low as 7.33 may represent short rides, discounts, or promotions. This variable may not follow a perfectly normal distribution, as the mean and median are not identical, and a slight skew may exist.

## Shapiro Wilk Test

By plotting the data within histograms, it can be observed that some graphs were slightly normalized, however, Shapiro Wilk test suggests that none of the variables follow a normal distribution, since the p-value on each column were lower than 0.05 as it shown in the picture below.

A screenshot of a computer program

Description automatically generated

Figure Shapiro-Wilk Test Results

The variables Rider Age, Duration, Distance and Cost reject the hypothesis of normality according to the Shapiro-Wilk test (p ≤ 0.05). Although some variables show apparent symmetry (Rider Age and Distance), all have characteristics that indicate deviations from the normal distribution, such as negative kurtosis (flatter distributions).

## Population Samples

### Anova Test

Due to the pricing strategies research coming from this dataset, the target variable will be Cost, and this will be applied to a regression model. Therefore 3 samples were taken in relation to rider satisfaction in order to stablish the confidence interval and Anova test.

Histograms were created for each one, which as a result do not present a normal distribution. In addition, the mean, standard deviation, standard error and the confidential interval for 95% were calculated as can be seen in the following image.

A screenshot of a computer

Description automatically generated

Figure Satisfaction Category Samples & Confidence Interval

The previous results suggests that there are no significant differences in the means between the High, Medium and Low categories, since the means are virtually equal, the confidence intervals overlap considerably the low standard error indicates that the means are precise estimates, supported by adequate sample sizes.

In order to support this theory (null hypothesis), Anova test has been applied, giving the following result.

A close up of a text

Description automatically generated

Figure Anova Test for Satisfaction Categories

The ANOVA results suggest that the cost does not significantly differ across satisfaction levels. Therefore, cost may not be a distinguishing factor between the High, Medium, and Low satisfaction groups.

### T-test

The variables Cost and City have been chosen to verify if the average costs are different depending on different cities for pricing adjustment.

Histograms were created for each one, which as a result do not present a normal distribution rather than Dublin City which is the only since the sample is smaller. In addition, the mean, standard deviation, standard error and the confidential interval for 95% were calculated as can be seen in the following image.

A screenshot of a computer screen

Description automatically generated

Figure City Samples & Confidence Interval

Usage of independent two sample (Z-score) has been selected to obtain two-tailed, because the samples have a number of observations greater than 100, (DataTap, 2024). Therefore T-test cannot be applied as it is for samples with observations less than 30. It should be emphasized that two independent tests were conducted to compare cities by continent, such as Berlin and New York or Berlin and London.

The null hypothesis for both cases is that there is no significant difference between the means.

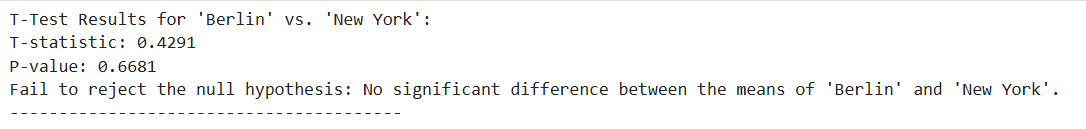
Berlin vs New York results:

Figure T-Test Results for Berlin and New York

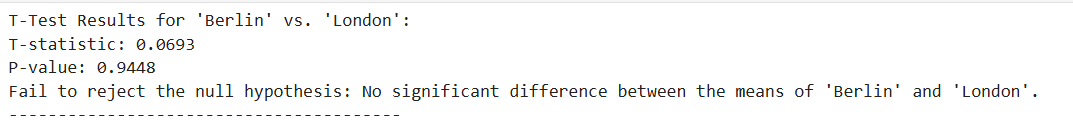
Berlin vs London results:

Figure T-Test Results for Berlin and London

As can be seen, the results suggest that there is no significant difference between the means. The p-value for both cases was greater than 0.05 which does not reject the hypothesis.

Furthermore, T-statistic of image1 and image 2 Indicates that the difference between the means is small or not large enough relative to the variability of the data.

### Covariance

A screenshot of a number

Description automatically generated

Figure Covariance Matrix

Duration and Cost have the highest positive covariance (289), suggesting that these variables are positively related: the longer the duration, the higher the cost. (Taylor, 2020).

Most of the other covariances are close to 0 (-4,1,2, -5), indicating that there is no strong linear relationship between these variables.

### Correlation

A screenshot of a graph

Description automatically generated

Figure Correlation Matrix

As for the correlation graph, the results confirm that Duration has a positive significance in relation to Cost.

## Regression

A linear regression model has been applied for the independent variable Duration and the independent variable Cost, which has yielded the following results.

A graph with blue dots

Description automatically generated

Figure Regression Model Results

The intercept could represent fixed costs that do not depend on duration, such as base rates or minimum usage costs. The coefficient suggests that duration has a positive, linear impact on cost.

The r2 value indicates that approximately 81.95% of the variation in Cost can be explained by Duration in this linear model. This is a relatively high value, suggesting that the model has a good explanatory performance. However, it is not perfect, which could indicate the influence of other factors not considered in the model.

The MSE indicates that there is a certain level of error, which is normal for this analysis.

# Machine Learning Model

The dataset provided includes variables related to rider satisfaction, demographics, trip characteristics, and environmental conditions. The target variable is Rider Satisfaction, which is categorical. This makes the dataset naturally suited for a supervised learning approach because:

The goal is to predict or analyse Rider Satisfaction based on the other variables.

Supervised learning models are designed to work with labelled data containing a context, (Delua, J. 2021).

**Pros:**

* Models such as Decision Trees or Random Forests can learn patterns in the data to predict Rider Satisfaction for the company to attract new riders.
* Feature importance can help identify which factors (e.g., distance, cost, weather) most influence satisfaction.
* Metrics like accuracy or F1-score provide clear feedback on model performance.

**Cons:**

* If the data collection process for satisfaction scores was inconsistent or biased, the model may inherit these issues.
* Complex models may be overfit to this relatively small dataset (900 samples).

Unsupervised learning is less suitable for this case, as it is typically used to:

* Discover hidden patterns or clusters (e.g., grouping riders by behaviour without predefined categories).
* Reduce dimensionality for exploratory purposes (e.g., PCA).

**Pros:**

* Can identify groupings or trends in the data that supervised methods might overlook.
* Useful when the target variable is unknown or unreliable.

**Cons:**

* Does not align with the goal of predicting or analysing satisfaction.
* The insights from clustering or dimensionality reduction may not directly translate into actionable findings.

To fulfil the company goals, it has been decided to work with Rider Satisfaction for customer satisfaction improvement and Cost variable for pricing prediction.

## Classifier Models

The following classification models were used for the different datasets such as original dataset, scaled dataset, PCA-reduced dataset, LDA-reduced dataset and in addition, different test sizes such as 30 and 15% were tested to verify how the accuracy results varied:

A screenshot of a computer code

Description automatically generated

Figure Classifier Models Integration

### 30% Split Test

Different tests were created to obtain the best results with different models, obtaining the following information.

**Original Dataset**

A screenshot of a graph

Description automatically generated

Figure Classifier Models Result for Original Dataset and 30% Split

**Scaled Dataset**

A screenshot of a graph

Description automatically generated

Figure Classifier Model Results for the Scaled Dataset and 30% Split

**PCA Dataset**

A screenshot of a graph

Description automatically generated

Figure Classifier Model Results for PCA Reduced Dataset and 30% Split

**LDA Dataset**

A screenshot of a graph

Description automatically generated

Figure Classifier Model Results for LDA Reduced Dataset and 30% Split

### 

### 15% Split Test

After having changed the train and test size of the model, simply to discover its overfitting and underfitting behaviour, the following results have been delivered.

**Original Dataset**

A screenshot of a graph

Description automatically generated

Figure Classifier Model Results for Original Dataset and 15% Split

**Scaled Dataset**

A screenshot of a graph

Description automatically generated

Figure Classifier Model Results for the Scaled Dataset and 15% Split

**PCA Reduced Dataset**

A screenshot of a graph

Description automatically generated

Figure Classifier Model Results for PCA Reduced Dataset and 15% Split

**LDA Reduced Dataset**

A screenshot of a graph

Description automatically generated

Figure Classifier Model Results for LDA Reduced Dataset and 15% Split

The graphs suggest that in some cases LDA at 15% of the test size performs better than at 30%. However, from this first stage it has been suggested that no model has a significantly high accuracy, F1-Scores are also low, indicating that the models are failing to correctly capture the complexity of the problem.

Furthermore, it could be observed from the correlation map that the Rider Satisfaction variable did not correlate with any of the independent variables, suggesting the poor performance of the models from the beginning.

### Hyperparameters Tunning and Cross Validation

Random Forest was used at 15% test on the original dataset due to the best results obtained, and KNN on the LDA-reduced dataset yielding the following results:

**Random Forest**

A screenshot of a report

Description automatically generated

Figure Random Forest Report

A blue squares with white text

Description automatically generated

Figure Random Forest Confusion Matrix

A screenshot of a computer code

Description automatically generated

Figure 10-Fold Cross-Validation for Random Forest

**K-Nearest Neighbours**

A screenshot of a report

Description automatically generated

Figure K-Nearest Neighbours Report

A diagram of confusion matrix

Description automatically generated

Figure K-Nearest Neighbours Confusion Matrix

A number on a white background

Description automatically generated

Figure 10-Fold Cross-Validation Results for KNN Model

### 

### Classification Model Conclusion

After a study of all the results, the model suggests that the available variables do not adequately explain variations in rider satisfaction. Even after feature engineering and the creation of additional variables, this weak relationship persists. No model achieves meaningful metrics in terms of accuracy or F1-Score, which is consistent with the lack of relationship between the variables and the target.

Techniques such as PCA and LDA, together with test size reduction, did not achieve a significant change in the results, suggesting that the data as a whole may not be adequate to predict the target, since the dataset may not capture crucial factors that actually affect user satisfaction, such as subjective elements (e.g., personal experience, expectations, specific events during the service).

## Regressor Models

In order to obtain an improvement for the cost adjustments, it was necessary to see how this was connected to other factors, in this case the independent variables. The following regression models were used for the dataset but with the target variable ‘Cost’ and 30% for split size:

A screenshot of a computer code

Description automatically generated

Figure Regression Models

The r2, MAE and MSE results for each of the models were as follows:

A screenshot of a computer

Description automatically generated

Figure Regression Model Results for Original Dataset and 30% Split

For the results shown, the best model was the Random Forest model due to a low error and a high R2 which means an excellent fit. However, Train R2 Score shows possible signs of overfitting due to the perfect R² in the training set. In order to mitigate this issue, the hyperparameters must be adjusted.

### Hyperparameters Tunning and Cross Validation

To find the best results, GridSearchCV was used, the following image present the results of this search.

A graph with blue dots

Description automatically generated

Figure Random Forest Results After Hyperparameters Tunning

For the 10-Folds Cross-Validation the results were as follows:

A screenshot of a computer code

Description automatically generated

Figure 10-Folds Cross-Validation Results

These results suggest that, after hyperparameter adjustment, the Random Forest model now has a good balance between fit and generalisability, at least within the training dataset evaluated with cross-validation.

### Regression Model Conclusion

The analysis carried out to optimise the company's costs focused on identifying the most relevant factors through feature engineering techniques and regression models. The correlation map revealed that the duration variable has a high relationship with cost, along with other important variables. This allowed for the construction of more accurate models.

Among the models evaluated, Random Forest stood out as the most effective due to its high R² and low errors (MAE and MSE), indicating an adequate fit for predicting costs. However, initial results showed possible signs of over-fitting in the training set, which was mitigated by hyperparameter adjustment using GridSearchCV and 10-partition cross-validation. After these optimisations, the model achieved an adequate balance between accuracy and generalisability.

In conclusion, the adjusted Random Forest model provides a reliable tool for predicting costs, which can significantly help the company to set more competitive prices aligned with its financial objectives.

# Conclusion

The analysis reveals that the variables available in the dataset have a weak relationship with user satisfaction, which limits the predictive capacity of the models evaluated. Despite applying techniques such as PCA and LDA, the prediction of satisfaction did not reach significant levels of accuracy, suggesting that external or subjective factors not captured by the dataset may be determinant.

During the analysis, inconsistencies were identified in the dataset that affected key variables such as cost, duration and distance. One notable example was a record where a user with a standard bicycle covered 23 km in 5 minutes, implying an unrealistic speed of more than 400 km/h. This type of error, along with others like it, underlines the need for more rigorous data cleaning and validation. In addition, the absence of clear rates for calculating costs per minute or per kilometre made it difficult to create a fitted variable that could correct for these inconsistencies and relate them mathematically to other variables in the dataset.

On the other hand, the Random Forest regression model proved to be effective in predicting costs, especially after hyperparameter adjustment and cross-validation. The findings highlight the importance of improving the quality and comprehensiveness of data collected for future analysis and emphasise the value of data cleaning and validation techniques in analytics projects. Finally, this work offers valuable lessons on the implementation of predictive models and strategies for optimising prices based on historical data.

# References

DataTab (2024). *t-Test, Chi-Square, ANOVA, Regression, Correlation...* [online] datatab.net. Available at: https://datatab.net/tutorial/unpaired-t-test.

Delua, J. (2021). *Supervised vs. unsupervised learning: What’s the difference? | IBM*. [online] www.ibm.com. Available at: <https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning>.

Jaadi, Z. (2024). *A Step by Step Explanation of Principal Component Analysis*. [online] Built In. Available at: https://builtin.com/data-science/step-step-explanation-principal-component-analysis.

Taylor, S. (2020). *Covariance*. [online] Corporate Finance Institute. Available at: https://corporatefinanceinstitute.com/resources/data-science/covariance/.

# Github

https://github.com/CCT-Dublin/ca2-AntonioGiambra