**Title: Predicting Taxi Fares in Miami with Random Forests Models**

**Abstract:** In my project, I explore how applying random forest models to predict taxi fares in Miami, Florida, utilizing an custom dataset that I developed through advanced SQL query techniques and methodologies by encompassing various factors such as location, time, and fare data, and refining it through the SQL query editor in R Studio. My analysis goal is to uncover patterns in fare distribution in Miami, FL and southern parts of southeast Florida by evaluating the effectiveness of machine learning in predicting transportation costs. By modifying the dataset of a random sample of 1000 observations and 8 variables utilizing SQL queries in R Studio to customize a dataset by filtering, joining, and restructuring data to enhance the accuracy and performance of the random forest model. Then with statistical formulas, to understand the model’s behavior by highlighting the strengths, weaknesses and limitations of random forest models. Allowing myself to oversee if the data allowed for better predictions in the transportation industry.

**Introduction:** In the era of big data, machine learning has become an essential tool for optimizing operations across various industries, including transportation amongst ridesharing programs such as taxi services. By utilizing predictive models with machine learning algorithms, I am able to take advantage of forecasting complex pattern trends in the data. In order to better understand the fare distribution in the city of Miami related to random forests in transportation costs utilizing prediction, my research paper focuses on applying random forest models to predict taxi fares in Miami. This allows for a more detailed geographical visualization of the actual predicted fares by zooming in on particular regions of southern Florida.

**Data Description:** The dataset utilized consists of 1000 observations and 8 variables in my statistical analysis of random taxi trip records from Miami, Florida. It includes fields such as medallion number, pickup datetime, geographic coordinates, trip duration, fare amount, and tip amount. The dataset was filter, clean and transformed by incorporating advanced SQL queries in R Studio by specifically identifying relevant columns, renaming variables for clarity, and filtering out non-positive fare values to ensure the accuracy and reliability of the analysis of the model utilized. My preparations in transforming the dataset was essential in order to provide the random forest model ability to learn from the data, in order, to provide an meaningful and accurate predictions of taxi fares by region.

**Methodology:** My approach involved first with data preparation, where the raw data was filtered from a sample to help transform the variables to include additional custom features or columns that needed to be merged together such as the hour, day of the week, and the month of the pickup times associated with taxi fares in Miami, Florida. These columns were of special interest to myself as I was able to identify pattern trends and variations in taxi fares in specific regions of the area that could influenced the outcome based on factors like rush hours and seasonal trends throughout the year.

My first approach was to allow the variable *Y*, to represent the taxi fare, and  to represent the independent variables such as the pickup location, time of day, day of the week of the taxi fares southern Florida. I was able to identify the relationship between the dependent variable Y and the independent variables  by incorporating it into an expression as follows:



Where f is the underlying function we aim to approximate using machine learning, and ϵ or epsilons represents the random error term of the model.

**Spatial Visualization:**

To gain a deeper understanding of the dataset and enhance the accuracy of fare predictions, I utilized the OpenStreetMap package within RStudio, a powerful tool for geographical visualization. This package enabled me to effectively map taxi data across various locations in Miami, FL, providing a visual context for the analysis. By integrating ggplot2, I applied a color schema that highlighted spatial distribution patterns through a scatter plot matrix. This approach allowed me to visualize and analyze the geographical spread of taxi fares, pinpointing areas of interest within the city. The scatter plot matrix not only revealed where the fares were most concentrated but also provided insights into how location-specific factors might influence fare variations. This comprehensive spatial analysis served as a crucial step in refining the model's predictive capabilities, offering a more nuanced exploration of fare distribution across Miami.

**Regression Tree Model:**

The regression tree model I utilized to explore the relationship between fare amount and the spatial-temporal features of the dataset encountered various problems due to the latitude and longitude being either a single node or to complex node populate a readable regression tree. The regression tree algorithm attempted to partition the data into subsets but I had to later redefine the nature of the dataset to make the predictors more homogeneous as possible by specifically concentrating on the target variable (taxi fare). After numerous attempts I was able to employ following equation to get a better grasp on the data patterns:



Where  is the actual fare,  , and the predicted fare for Miami, FL regions  ​, and *M* is the number of regions (or leaves) in the regression tree model.

**Random Forest Model:**

My focus of the analysis was on building, transforming and optimizing the highest accuracy of the random forest model by allowing the model to learn different ways to construct a specific method that constructed multiple decision trees during training and outputs that allowed the mean of the prediction of the individual trees more accuracy. Below you will find the random forest algorithm that I utilized for my analysis could be represented as the following:

A black background with a black square

Description automatically generated with medium confidence

Where  is the prediction for the new data points on the geographical map of southern Florida with *X, B* is the number of trees, and  is represented by the prediction of the *b-th* regression tree model. Looking closely you can see how each tree is trained on a bootstrap sample of the dataset, and a random subset of features is considered for each split in the tree was to enormous at first so I had to reduce it either further to get a better representation of the dataset.

This involved the model to be fine-tuned before I adjusted the parameters to incorporate the key number of trees *(ntree)* and the sample size *(sampsize)* to be represented in a more fashionable representations for better clarity. The parameters I choose to reduce the dataset were necessary for better model performance and accuracy to avoid overfitting.

**Results:** The results of the analysis are presented in terms of both spatial patterns and temporal trends of the latitude being less than 30.9819 to split between the latitude of 29.8614. Further narrowing down the geographical location of the highest predicted fares in between the regions of 51.78 latitude and 44.03 longitude of southern Florida. To clarify the accuracy of the model I utilized various heatmaps within ggplot2 to map out with an color gradual schema of rectangle colored boxes to focus on the predicted fares across Miami. This allow and revealed a significant difference in fare distributions in different areas of the region based on highest results of income generation for taxi drivers. The higher fares were often observed in regions with higher demand for taxis associated with areas more focused on downtown Miami, local beaches and attractive tourist hotspots were transportation was limited for travels or individuals traveling to work. These results were consistent with my hypothesis that fare amounts vary based on location and time of day played a pivotal role compared to rural areas away from the city.

**Model Evaluation:**

The performance of the random forest model was evaluated several times to ensure the accuracy of the mean squared error (MSE) to make sure the model was performing accurately, by incorporating statistical analysis with this formula, which was calculated with this formula:



Where  , In conclusion, this study effectively demonstrates the application of random forests for predicting taxi fares in Miami. By leveraging spatial and temporal data, the model offers a nuanced understanding of fare distributions across the city, providing valuable insights for taxi service providers, urban planners, and policymakers. While the random forest model is adept at capturing complex patterns, its performance reveals areas for improvement, particularly in off-peak times. The study highlights the potential benefits of incorporating additional data sources, such as traffic conditions or weather data, to enhance prediction accuracy.

The findings underscore the model's strengths and limitations, contributing to the growing body of knowledge on data-driven approaches in urban transportation planning. This emphasizes the importance of integrating statistical methods and machine learning in modern predictive analytics, while also acknowledging the need for careful consideration of model applicability and potential enhancements.

In conclusion, this study effectively demonstrates the application of random forests for predicting taxi fares in Miami. By leveraging spatial and temporal data, the model offers a nuanced understanding of fare distributions across the city, providing valuable insights for taxi service providers, urban planners, and policymakers. While the random forest model is adept at capturing complex patterns, its performance reveals areas for improvement, particularly in off-peak times. The study highlights the potential benefits of incorporating additional data sources, such as traffic conditions or weather data, to enhance prediction accuracy.

The findings underscore the model's strengths and limitations, contributing to the growing body of knowledge on data-driven approaches in urban transportation planning. This emphasizes the importance of integrating statistical methods and machine learning in modern predictive analytics, while also acknowledging the need for careful consideration of model applicability and potential enhancements.

**Discussion:**

In conclusion, this study effectively demonstrates the application of random forests for predicting taxi fares in Miami. By leveraging spatial and temporal data, the model offers a nuanced understanding of fare distributions across the city, providing valuable insights for taxi service providers, urban planners, and policymakers. While the random forest model is adept at capturing complex patterns, its performance reveals areas for improvement, particularly in off-peak times. The study highlights the potential benefits of incorporating additional data sources, such as traffic conditions or weather data, to enhance prediction accuracy.

The findings underscore the model's strengths and limitations, contributing to the growing body of knowledge on data-driven approaches in urban transportation planning. This emphasizes the importance of integrating statistical methods and machine learning in modern predictive analytics, while also acknowledging the need for careful consideration of model applicability and potential enhancements.

**Conclusion:** In conclusion, this study effectively demonstrates the application of random forests for predicting taxi fares in Miami. By leveraging spatial and temporal data, the model offers a nuanced understanding of fare distributions across the city, providing valuable insights for taxi service providers, urban planners, and policymakers. While the random forest model is adept at capturing complex patterns, its performance reveals areas for improvement, particularly in off-peak times. The study highlights the potential benefits of incorporating additional data sources, such as traffic conditions or weather data, to enhance prediction accuracy. The findings underscore the model's strengths and limitations, contributing to the growing body of knowledge on data-driven approaches in urban transportation planning. This emphasizes the importance of integrating statistical methods and machine learning in modern predictive analytics, while also acknowledging the need for careful consideration of model applicability and potential enhancements.