Random Forest Review

1. Random Forest:

Categorial Features: *member_casual , season*
Contínuos Features: *TMAX*
Target Variable: *rideable_type*
NaN Values: dropped all the rows with null values

Random Forest Accuracy: 0.64
Random Forest Confusion Matrix: [[1650972 10486]
 [937518 11376]]

The Random Forest model achieves an accuracy of 64%, indicating moderate effectiveness in predicting the correct rideable_type. However, the confusion matrix reveals a significant imbalance in predictive performance: while true negatives are high at 1,650,972, suggesting good specificity, the model suffers from a high number of false negatives (937,518), indicating poor sensitivity or a strong bias toward the more prevalent class. This results in only 11,376 true positives, which is concerning for applications needing reliable identification of the positive class. The low false positives (10,486) suggest that the model is conservative, potentially at the cost of missing many true positive cases.

Categorial Features: *member_casual , season*
Contínuos Features: *TMAX*
Target Variable: *rideable_type*
NaN Values: Handled NaN Values.

Random Forest Accuracy: 0.53
Random Forest Confusion Matrix: [[749533 916209]
 [674585 1075940]]

The Random Forest model shows an accuracy of 53%, which is lower than might be desirable for robust classification tasks. The confusion matrix indicates a nearly balanced but still challenging scenario with considerable false positives

(916,209) and false negatives (674,585), alongside true positives (1,075,940) and true negatives (749,533). This performance suggests that handling null values by imputation rather than dropping them has not necessarily improved the model's predictive accuracy or the balance between sensitivity and specificity. The model's ability to correctly classify the positive class has improved in terms of true positives but at the cost of increasing false positives, highlighting a trade-off between detecting more positives and making more errors in classifying negatives.

Categorial Features: *member_casual , day_of_week*
Contínuos Features: *TMAX*
Target Variable: *rideable_type*

Random Forest Accuracy: 0.54
Random Forest Confusion Matrix: [[843729 822013]
 [746061 1004464]]

The Random Forest model's performance with a new feature set achieves an accuracy of 54%, a slight improvement from previous models. The confusion matrix reveals a more balanced result between true positives (1,004,464) and true negatives (843,729), but also high numbers of false positives (822,013) and false negatives (746,061). This indicates that while the model has become slightly better at identifying both classes, it still struggles with a significant error rate in both predicting false outcomes. The introduction of new features appears to have moderately enhanced the model's ability to detect positive cases

Categorial Features: *member_casual , day_of_week, season*
Contínuos Features: *TMAX, Elevation_Change*
Target Variable: *rideable_type*

Random Forest Accuracy: 0.56
Random Forest Confusion Matrix: [[925987 739755]
 [774323 976202]]

The Random Forest model with a new feature set yields an accuracy of 56%, showing a modest improvement over previous versions. The confusion matrix presents a mixed scenario with 925,987 true positives and 976,202 true negatives, balanced against high false positives (739,755) and false negatives

(774,323). This result indicates a slight improvement in model performance, capturing more true cases but still suffering from a considerable number of incorrect predictions.

Feature importances derived from a Random Forest Classifier model:

Feature	Importance
member_casual_member	0.468636
member_casual_casual	0.262494
season_Fall	0.154286
season_Spring	0.046343
season_Summer	0.045321
season_Winter	0.003168
day_of_week_Friday	0.002681
TMIN	0.001925
TMAX	0.001881
Elevation_Change	0.001878
trip_duration	0.001854
Distance	0.001789
day_of_week_Wednesday	0.001770
day_of_week_Tuesday	0.001581
day_of_week_Monday	0.001404
day_of_week_Saturday	0.001390
day_of_week_Sunday	0.001071
day_of_week_Thursday	0.000530

The most influential features are member_casual_member and member_casual_casual, with importance scores of approximately 0.469 and 0.262 respectively. This indicates that the type of membership (casual vs. member) is highly predictive of the target variable, which in this context could be the type of ride or user behavior in a bike-sharing dataset. The much higher score for members suggests that whether a user is a registered member is a strong indicator in the model.

Seasonal variables also play a significant role, especially season_Fall, which has an importance score of about 0.154. The importance decreases for other seasons like Spring, Summer, and especially Winter, which has the lowest score among the seasons. This pattern might reflect seasonal usage trends specific to the dataset’s geographical and climatic context.

The features representing days of the week have relatively low importance, all scoring below 0.003, with Friday being the highest among them. This suggests that while there is some variation in ride patterns across the week, it's not as significant as the type of user or the season.

Among the continuous features like Elevation_Change, Distance, trip_duration, TMAX, and TMIN, all have very low importance scores ranging from about 0.0019 to 0.0018. This surprisingly low importance could indicate that these conditions, while potentially impactful to a ride’s difficulty or duration, do not vary significantly with the type of rideable or are overshadowed by categorical data like member_casual and season.