Annual Membership (Business Plan)

ISSUE / PROBLEM

Citibike seeks to improve Business plan for 2016 and answer the following question:

What's likely to make people become annual membership?

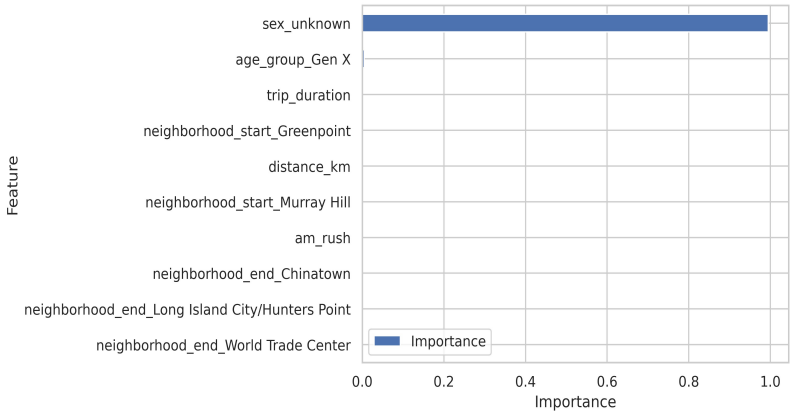
RESPONSE

Since the variable we are seeking to predict is categorical, the team could build either a logistic regression or a tree-based machine learning model or a xgboost.

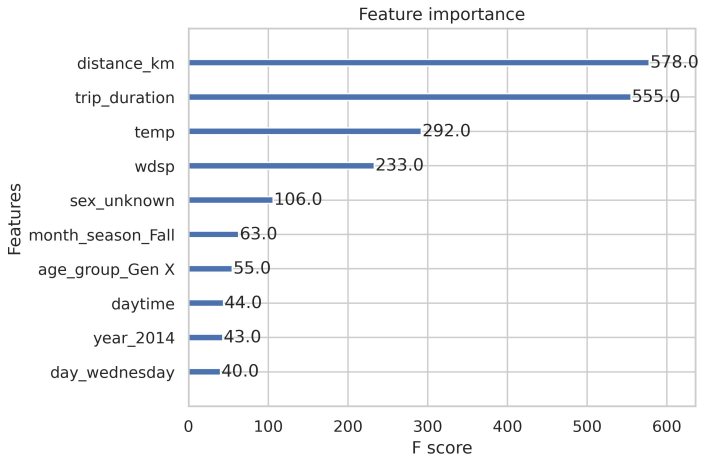
The models perform 100% the random forest and xgboost model.

IMPACT

This model helps predict whether people will become an annual member which factors are most influential. These insights can help make decisions to improve demand side.



Barplot above shows the most relevant variables with random forest: *sex_unknown', age_Gen_X', 'trip_duration', 'neighborhood_departure' and 'distance_km'.*



In theXG boost model above, 'saistance_km', 'number_project', 'trip_duration", 'temp', 'wdsp', and `sex_lunknown` have the highest importance. These variables are most helpful in predicting the outcome variable, `annual membership`.

INSIGHTS/NEXT STEPS

- Establish initiatives for safe cycling, including safety training and education, distribution of safety equipment, and related endeavors.
- Foster community partnerships and programs, such as bike ambassadors, youth and family engagement, grants for low-income individuals, etc..
- Expand station or location within the top departure and destination neighborhoods particularly as displayed in the feature importance results. For instance, consider the departure stations in Greenpoint and Murray Hill, as well as the destination stations in Chinatown and the World Trade Center. Additionally, conduct an in-depth exploration of the unique attributes that define these prominent locations.
- The growth observed in 2015 demonstrates a clear seasonality pattern, characterized by a gradual rise during the summer months, reaching its peak in September during the fall, and subsequently experiencing a gradual decline. This trend allows us to concentrate on medium to large maintenance activities during the slower seasons. These maintenance activities could include workshop repairs, strategic planning, or acquiring new assets. The execution of these maintenance efforts can be initiated after the winter period, just before the onset of the summer season when the number of riders is expected to rise. This is an opportune time to implement on-the-spot repairs as needed.

- Allocate resources for repair and maintenance on-site, with a particular focus on quality checks in top trip areas during off-peak hours and particularly warm season.
- Hire both seasonal and permanent staff members who can actively engage with communities and contribute to the organization's objectives.
- Implement marketing and promotional awareness campaigns, such as #Bike4Youth and bike health events, as well as prominently displaying the NYCHA residents rate $5 per month at stations.

Feaiture selection and engineering:

- Scenario of possibilities to understand this model is required. By considering the data from the four models with the complete dataset, we observe a 100% accuracy across decision tree, random forest, and XGBoost models. Building upon the insights shared in part 2, we can further delve into the analysis of trip distances categorized as short, medium, and long distances. Regarding trip durations, we have introduced additional columns to account for overtime. In the future, we might consider incorporating another duration metric that reflects the permissible time during the day, potentially leading to the removal of the original trip duration column. Furthermore, we can consider creating columns that categorize areas based on factors such as subway access, residential nature, commercial activity, recreational facilities, Instagram-worthy spots, and tourist attractions. Also, we can distinct the bike id into e-bike and reg-bike columns. Moreover, leveraging Pareto data, we can stratify trips into quartiles: the top 25% as high trips, the middle 50% as medium trips, the subsequent 80% as low trips, and any beyond the 80% threshold as developmental trips.
- As also mentioned earlier, the availability of user IDs allows us to categorize bikers into distinct groups such as platinum, frequent, high, medium, and low bikers. This approach underscores the importance of user ID tracking, enabling us to determine the precise number of registered members who actively engage with the bike-sharing service. This, in turn, aids us in determining the optimal allocation of bike investments based on the observed demand patterns. Furthermore, we can conduct surveys to gain insights by sampling from the members, determining whether it is the demand or supply side that requires attention and adjustments. This approach unlocks further potential for feature refinement.