Machine Learning Model Outcomes

Executive summary report for the New York City Citibike

Prepared by Wahyu Ardhitama

Overview

New York City Citibike would like to establish a business plan for 2016 to build an analysis from customer insights to grow the business and create a machine learning model to predict whether more annual membership utilize the bikes.

Problem

The challenge we encountered was developing a business plan without knowing the exact user and revenue, so we focus on total trip and trip duration in the demand side. We also considered ethical consequences of building this model, particularly regarding the errors that we will encounter when the model is built.

Solution

We built two different models and compared results and parameters. Despite hypertuning and boosting, the **outcome keep showing a perfect 100%**. We would recommend using this model by combining feature importance.

Details

Behind the data

- Our assumption was that as this a short ride which a mides id transportaion fpr people who aretoo far walking but to near using bus. Location, month and time of day may have a strong enough relationship with total trips then we could accurately predict annial membership from just these variables.
- As we built our models and performed the testing, it became clear that there was not as strong a correlation as anticipated; but rather a passable one.

Models	AUC	Prec	f1 Score	Recall	Acc
Random forest	100%	100%	100%	100%	100%
XG boost	100%	100%	100%	100%	100%

Future model suggestions

- Remove “year” 2014 as a feature
- Add distict user id and bike id to identify their types
- Predict supply side
- Cluster with K-means and analyze the clusters to derive insights from the data

Results Summary

The resulting algorithm is usable to predict annual members, but it is a model that we still have to validate as the models shows 100% in its matrix.

Next Steps

- Create a new feature engineering and conduct analysis on the supply side
- Share the results and recommendation that the model could be used as an indicator, as long as there is an understanding of its limits on correct prediction.