PROJECT MILESTONE 6

PACE: Plan Stage

- What am I trying to solve or accomplish?

I'm trying to develop machine learning model that predicts whether a taxi cab rider will be a generous tipper.

- What resources do I find myself using as I complete this stage?

- taxi trip data

- taxi company means of predictions

- Jypiter notebook

- Do I have any ethical considerations at this stage?

Drivers who didn't receive tips will probably be upset that the app told them a customer would leave a tip. If it happened often, drivers might not trust the app. Drivers are unlikely to pick up people who are predicted to not leave tips. Customers will have difficulty finding a taxi that will pick them up, and might get angry at the taxi company. Even when the model is correct, people who can't afford to tip will find it more difficult to get taxis, which limits the accessibility of taxi service to those who pay extra. Therefore, initial modeling objective (predicting non-tippers) was rejected, and it was decided to predict "generous" tippers—those who tip ≥ 20%.

- Is my data reliable?

Yes, the provided data is reliable after doing some EDA practice.

- What data do I need/would like to see in a perfect world to answer this question?

I would like to have a history of every customer's tipping habits, which will help me to predict future tips.

- What data do I have/can I get?

I have a complete dataset with a lot of detailed information about every ride in 2017.

I can get behavioral history for each customer, so I could know how much they tipped on previous taxi rides.

- What metric should I use to evaluate success of my business/organizational objective? Why?

I should use metrics like times, dates, and locations of both pick-ups and drop-offs, estimated fares, mean duration and distance of rides.

Those metrics should be strong predictors whether a rider will give generous tip >=20% or not, because the longer the trip is, the probability of getting generous tips is higher. Also, the time of day (rush-hour), pick-up, drop-off locations all play significant role in tipping.

PACE: Analyze Stage

- Revisit “What am I trying to solve? “Does it still work? Does the plan need revising?

The initial objective was to predict if a rider will leave a tip. But, from ethical considerations, that objective was changed to predict if a rider will leave a tip >=20% of the taxi fare – generous rider, which is ethically OK.

- Does the data break the assumptions of the model? Is that ok, or unacceptable?

Setting the value of 20% over the fare amount showed that there is almost balance between lower and generous tippers (47/53%), so this is acceptable.

- Why did I select the X variables I did?

I should use times, dates, and locations of both pick-ups and drop-offs, estimated fares, mean duration and distance of rides. Because tips are allowed only for payments with credit card, the dataset should be filtered to only rides with credit card payments.
Those metrics should be strong predictors whether a rider will give generous tip >=20% or not, because the longer the trip is, the probability of getting generous tips is higher. Also, the time of day (rush-hour), pick-up, drop-off locations all play significant role in tipping.

- What are some purposes of EDA before constructing a model?

1. Data Understanding

2. Feature Engineering

3. Data Cleaning

4. Model Selection

- What has the EDA told me?

I understood that those tips are allowed only with credit card payments, so I filtered the dataset with only credit card payments (15.265 out of 22.699 total records).

Further I considered time of day (am and pm rush hours, daytime, nighttime) to be strong predictors, so I created new variables based on pick-up and drop-off datetimes.

I also created weekday and month columns considering that they could be predictors, too.

- What resources do I find myrself using as I complete this stage?

- 'Taxi_Trip_Data.csv'

- 'taxi_preds_means.csv'

- Jupiter notebook

PACE: Construct Stage

- Do I notice anything odd? Is it a problem? Can it be fixed? If so, how?

I noticed that Random Forest model performed better than XGBoost model, which is rare. It is not a problem, but it is worthy to make additional parameter tuning to compare the results between the two different models.

- Which independent variables did I choose for the model, and why?

I've chosen 15 independent variables which I considered the most predictive for the model objective.

After the analysis of the feature importance, maybe I would drop some of them in the next iteration of the models.

- How well does the model fit the data? What is my model's validation score?

Both models (RF and XGBoost) fit very well in the data, with RF F1-score of 0.7136 which is better than XGB F1-score of 0.6977.

- Can I improve it? Is there anything I would change about the model?

Maybe I would feature additional categorical variable 'trip_distance_type' with three possible values: 'short', 'medium' and 'far' because it can also be a predictor for tipping.

Also, I would try with additional parameters tuning on both models.

- What resources do I find myself using as I complete this stage?

- joined dataset

- Jypiter notebook

PACE: Execute Stage

- What key insights emerged from the model(s)? Can I explain my model?

The champion RF model is usable to predict riders who might be generous tippers, with reasonably strong precision, recall, F_1, and overall accuracy scores.

- What are the criteria for model selection?

The value of F_1 score is the criteria for RF model selection, because both false positives and false negatives should be kept minimal.

- Does my model make sense? Are my final results acceptable?

The RF model confirmed that variables with most predictive power to generous tips are 'predicted_fare', 'mean_duration', 'passenger_count' and 'mean_distance'. Also, the RF model showed that 'VendorID' variable has the biggest predictive power, which requires further examinations.

- Do I think the model could be improved? Why or why not? How?

The model could be improved with additional information about tipping behavior of passengers.

It would also be valuable to have accurate tip values for customers who pay with cash. It would be helpful to have a lot more data. With enough data, we could create a unique feature for each pick-up/drop-off combination.

- Were there any features that were not important at all? What if I take them out?

The feature importances attribute of the best estimator object showed that only 4 features were important and the others could be removed. The models should be reevaluated with these predictors, as well as to add some additional engineered features.

- What business/organizational recommendations do I propose based on the models built?

As a next step, I can consult the taxi company to share the model results and recommend that the model could be used as an indicator of tip amount. However, additional data would be needed to realize significant improvement to the model.

- Given what I know about the data and the models I were using, what other questions could I address for the stakeholders?

I would ask for additional information on customers tip behavior, including cash tips, as well as more data to predict on.

- What resources do I find my using as I complete this stage?

- Jupiter notebook

- Is my model ethical?

By dropping the initial classification on tippes and non-tippers, by setting a threshold on 20%, it can be considered as ethical model.

- When my model makes a mistake, what is happening? How does that translate to my use case?

When the model false predicts the outcome, there is two possible scenarios:

1. Model makes false positive predictions, meaning that the driver is expecting a generous tip when the rider actually won't give a tip.

2. Model makes false negative predictions, meaning that the driver is surprised with a generous tip when he is not expecting.

Unfortunately for the drivers, the model makes more false positives (almost 20% of all predictions) than false negatives (11% of all predictions). But, still the model is performing very well with 69% accurate predictions which is almost 20% above the random guess.