BASKET ANALYSIS

FOR INSTACART

TING SIT

JUNE 2020

OBJECTIVE

- Predict user next purchase products combinations
- Benefits:
  - Customize product pages for users to increase conversion
  - Improve consumer experience with more relevant products to browse
  - Help retailers to plan for inventory based on expected product demand

Model Evaluation requirements:

- F1 score

MODEL CONSIDERATION

- Classification problem: reordered by user_id and product_id
- 4 models: Xgboost, Random Forest, Lightbgm, Gradient Boosting
- Features:
  - Number of times reordered in total
  - Number of times ordered for last three purchase
  - Product exist in the last purchase (category)
  - Average ‘days lag’ among between each purchase at user and department level
  - Frequency (%) of a product appear among all orders at user level

RESULTS

Model	F1***
Naïve*	0.31180
Random Forest**	0.36025
Xgboost**	0.36049
Lightgbm	0.36354
Gradient Boosting	0.36244

**Naïve model assume users purchased the same set of item as last order*
***Both Random Forest and Xgboost dropped 'day lag' feature*
**** F1 scores from all model used ~0.17 as the probability threshold.*

- Light Gradient Boosting has the highest F1
  - Recommend as the winning model

CONCLUSION

1. Light Gradient Boosting has the highest F1 score and is the winning model
2. Model generally has a recall rate of 0.5 (i.e. half of the qualified product purchased could not be identified.) It could be driven by model limitations on predicting new product purchases or products substitute within departments.
3. Lowering binary probability threshold is one option (as used in the solution), but that come with tradeoff from lower precision which clients need to factor into consideration.

NEXT STEPS

- Further improvement in F1 score could be achieved through new features identification or 2 step models to engineer on the F1 score.
  - The tradeoff will be potential overfitting.
- Align expectations with clients on the F1 level that is "good enough"
- Obtaining new feature information such as product prices/promotion could be useful, though that will increase the cost of computations since prices across different retailers could vary for the same item.