**WHAT WILL YOU EAT THIS WEEK?**

**PREDICTING INSTACART PURCHASES**

**Summary Paper**

**Objective:** Objective of the project is to develop a model that predicts which previously purchased products will be in a user’s next order.  For this, choose Random Forest Classifier, AdaBoost Classifier (Adaptive Boosting) and LightGBM as a prediction models.

**Data Source:** Source of data is open-sourced data from the Instacart Kaggle competition from 2017 for this project. (Source:<https://www.kaggle.com/competitions/instacart-market-basket-analysis/data>)

**Data Preparation:** The Dataset of Instacart is anonymized and has as samples of over 3 million grocery orders from more than 200,000 users, containing over 30 million records altogether. The dataset consists of 6 csv files that were merged to facilitate analysis.  The relationship diagram below illustrates the dataset files for customer orders.

A picture containing calendar

Description automatically generated

**Dataset Characteristics:**

What percentage of customers order the same items again? 59% of the products were reordered products, suggesting that users tend to reorder the same items at every purchase. Produce is the aisle where reorders are most likely with Banana on the top.

When do customer’s order: The day of the week has a direct effect on when orders are placed. Most orders are on days 0 and 1. Unfortunately there is no information regarding which values represent which day, but one would assume that this is the weekend. There is also an effect of hour of day on order volume. Most orders are between 10 am to 6 pm.  Customers seem to order more often after exactly 1 week or once in a month.

Most often reordered product: Organic products have the highest probability of being reordered followed by product names containing: chocolate, cheese, free (as in, gluten free for example), chicken, original, sauce, cream, yogurt, and mix.

Department with Highest reordered product:  All 3 models indicate that products from the household department are the highest recorded products.

**Analysis:**

The following predictor variables were used for the analysis:

* Categorical: 'aisle\_id', 'department\_id'
* Numerical: 'user\_id', 'order\_number', 'order\_dow', 'order\_hour\_of\_day', 'days\_since\_prior\_order', 'product\_id'

A random subset of 100,000 records from the merged “prior’ and “train” data was taken for the analysis. 80% of this subset is used for training and 20% for testing. Categorical variables were encoded using OneHotEncoder and numerical variables were transformed using MinMaxScaler.

Hyperparameters were tuned using GridSearchCV with 5-fold cross validation and roc\_auc used for scoring. The tuned models were used to make predictions on the test dataset with a comparison of model performance indicated in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Random Forest** | **AdaBoost** | **Light GBM** |
| **ROC\_AUC score** | 0.774 | 0.767 | 0.781 |
| **Accuracy score** | 0.719 | 0.717 | 0.668 |
| **F1 score** | 0.783 | 0.78 | 0.78 |

**Conclusion:**

Model evaluation results indicate best performance from the Light GBM classifier (according to the roc\_auc score); however, an assessment of learning curves suggests that the model is overfitting (training scores higher than cross-validation scores). Overfitting might be due to too much flexibility in the model (e.g., number of features included) or training for too long.  Despite relatively good performance, it is evident that improvements can be made perhaps through the inclusion of additional data (although computationally expensive); tuning additional hyperparameters; or an alternative modelling approach such as neural networks (where it would be no longer necessary to feed all the data into computer memory). Future work may wish to investigate feature engineering such as the inclusion of product keywords as a variable of interest. This may be valuable from a marketing perspective to recommend products to users.