WiDS ‘22 - ‘23 Final Documentation
Project UID: 8
Mentor: Rushi Chavda

Member Name	Roll Number	Email-Id
Raushan Kumar	200070068	200070068@iitb.ac.in

Introduction To Problem Statement:

The Problem statement is to predict which grocery products a customer will purchase again and when. The goal of this project is to predict grocery **reorders**: given a user’s purchase history (a set of orders, and the products purchased within each order), which of their previously purchased products will they repurchase in their next order?

The sequential and time-based nature of the problem also makes it interesting: how do we take the time since a user last purchased an item into account? Do users have specific purchase patterns, and do they buy different kinds of items at different times of the day?

Proposed Solution:

- Loaded all provided data and changed the data type of columns as needed to avoid errors.
- Did Data Pre-processing(i.e. Replacing null values with string s wherever needed) to clean data even if the data was already clean(because even without cleaning, it was working
- Analyzed the data by printing data just to get an insight as to what is going on with the data.
- Computed product frequency by creating a dataframe that shows the reorder rate.

- Added order dataframe columns to prior dataframe columns.
- Created "computing user frequency", in which we calculated Average days between two orders, and total number of orders made by a user.
- Now from priors we have added, total items bought by a user, all products the user bought, and how many distinct items users bought.
- Computed (user*product).
- Used LightGBM by asking mentors.

Methodology & Progress (Mention the work done week-wise):

Week-1: Going through the revision of basics of Machine learning, Pandas, Numpy and Meeting with Mentor

- Machine Learning tutorials: https://www.youtube.com/watch?v=LcWFedjaR4Q&list=PLfFghEzKVmjvII5ZcBnFWQOUjtUVdDnmo

Week-2: Going through the revision of basics of Pandas, Numpy, Matplotlib, Seaborn etc. and Instacart market basket analysis

- Pandas, Numpy, Matplotlib and Seaborn tutorial revision: https://www.youtube.com/watch?v=qAgyemeRhTw&list=PLfFghEzKVmjsgZPk2zxRRRLXCT0QyN495
- Instacart market Basket Analysis(Read ¼ the part of this analysis): https://medium.com/kaggle-blog/instacart-market-basket-analysis-feda2700cded

Week-3: Completed the Instacart market Basket analysis and Started reading and asking mentor about how to use XGBoost and LightGBM.

- Instacart market Basket Analysis(Read whole part of this analysis): https://medium.com/kaggle-blog/instacart-market-basket-analysis-feda2700cded
- Started implementing with the help of Mentor.

Week-4: Implementation

- Implementation using pandas, lightGBM and verification with mentor
- Making of report

Results:

Google Drive Link Containing all data files and .ipynb file: https://drive.google.com/drive/folders/1Clx17YvmhzzSwlX3Ah_3eusfpbXK5-gv?usp=share_link

Learning Value

- Revised Basics of Machine Learning even if I had no use of the technology in this project like NLTK library, NLP etc.
- Learnt how to build a movie recommendation system using **Tfidfvectorizer** and **cos_similarity.**
- Revised Pandas and numpy, Matplotlib, seaborn etc.
- Gone through the documentation of LightGBM and XGBoost and learnt at least the high-level-idea of these two models.
- Read the Instacart Market Basket Analysis documents shared by mentor.

Tech-stack Used:

Pandas, Numpy, LightGBM, Google Colab etc

Suggestions for others:

The only suggestions for others is to start from the basics and try not to grab everything at one go instead do implementations at the same time while going through the topics.

References and Citations:

- https://www.youtube.com/watch?v=A_78fGgQMjM
- https://www.youtube.com/watch?v=7rEagFH9tQg&ab_channel=Siddhardhan
- https://medium.com/kaggle-blog/instacart-market-basket-analysis-feda2700cded
- https://xgboost.readthedocs.io/en/stable/
- https://lightgbm.readthedocs.io/en/v3.3.2/index.html

- https://www.kaggle.com/competitions/instacart-market-basket-analysis/data
- https://developers.google.com/machine-learning/decision-forests/intro-to-gbdt