Recommendation System

Initial Review

Problem Statement:

1. Which users are going to buy something?
2. What items will user buy?

Data Analysis:

1) A large dataset of user sessions is provided.Each session contained user item click history including timestamp, item id and item category for all clicked items.
2) Dataset of buy events is also provided.Each buy event contained a corresponding session id, bought item id, time stamp, quantity bought and price.
3) In order to emphasize the differences between various sessions with and without buying events, we investigated and plot buy/click ratio of the sessions as the functions of various features.
4) From the graphs plotted various insights about data is developed.
    - Visiting a e-commerce site during mid-day leads to a purchase several times more than the night hours.
    - Purchase during a session has high probability on weekends than on working day.
    - High Numbers of items clicked during a session leads to a higher chance of purchase.
5) As only 5.5% of user-sessions have at least one buy event , therefore data contains class imbalance problem. As conventional classifiers are biased towards majority class, therefore will result in poor accuracy.

Approaches to solve class imbalance problem -

1. Data Level Approach - We can use oversampling and undersampling. Undersampling reduces the size of training dataset and therefore reduces training time and memory usage.
2. Cost-Sensitive Learning Approach - Give more weight to impose high cost of misclassification on the class of interest.

Various approaches to model the data -

1. Two stage classifier - In stage one(Purchase detection) use entire training set to predict whether user will buy some product and in stage two (purchase item detection) use only sessions with bought items in the learning phase to predict the items user will buy.Output will be the probabilities.
2. Single Classifier - For a given click instance , ignore the session Ids and predict whether a user will buy the currently clicked items or not. Once we get the prediction result for any clicked instances, we can tell that a session with any positive predicted click will end in a purchase.
3. As main problem with most of the models is to do feature engineering i.e extract information features from the original data. Therefore we can also use neural network modelling to let the model learn the sessions features from the sequence of raw click events in the session automatically. Bidirectional Recurrent Neural Network (BiRNN) based on LSTM can model click events directly P(y1,y2…..yn|x1,x2…..xk) for a session of click events {x1,x2,.....xn}.

List of Research Paper Referred -

- Ensemble learning with categorical features.
- Two-Stage Approach to Item Recommendation from User Sessions.
- An ensemble approach for multi-label classification of item click sequences.
- Predicting User Purchase in E-commerce by Comprehensive Feature Engineering and Decision Boundary Focused Under-Sampling.
- Linear and Non-Linear Models for Purchase Prediction.
- Neural Modeling of Buying Behaviour for E-Commerce from Clicking Patterns.

Note: All research papers mentioned in document is present in github repository.