Report of Walmart Kaggle project

Yunsong Zhang

1. Data preprocessing

As a first attempt, I only focus on the columns "Weekday", "VisitNumber", "DepartmentDescription" from the data, and also " Triptype" from the training data set. Nan values are replaced by "UNKNOWN"s for the variable "DepartmentDescription" and -1 for "FinelineNumber" and "Upc" columns.

It turns out there are 69 different Departments in the data, including the "UNKNOWN". I turn my data into an $n \times 70$ matrix, with the 70 features as weekday (encoded by 0 to 6) and the Scan Count for each Department in each visit.

2. Learning Algorithms

I learned to train the data with XGBoost and Extra-Tree algorithms. And it turns out both algorithms perform well with the test data prediction, while XGBoost beats Extra-Tree a lot in my first attempts with only weekday and department variables.

3. Further improvements

To further improve my results, I continue to consider data in the “FinelineNumber” column. However, since there are more than 5000 fine line numbers, I did dimension reduction here with PCA. The raw features are weekday, scan counts for each department and for each fine line in every visit. By performing a PCA, I have extracted 472 features, which are capable to explain more than 90% of the variance. Training the new data with XGBoost, I can improve the final prediction results.

4 Results

To summarize, my best submission with Extra-Tree scores 2.23094 in the public leaderboard and 2.22559 in the private leaderboard. While the XGBoost algorithm gets me a best score of 0.99026 in the public leaderboard, and 0.97209 in the private leaderboard. And the improving efforts with more features and PCA have scored me 0.83726 in the public leaderboard and 0.82692 in the private leaderboard.

As for parameter tuning, I did attempt to tune the parameters for both XGBoost and Extra-Tree with GridSearchCV function in the python scikit-learn package. However, it turns out to be extremely computationally expensive. I chose not to finish it, for this is merely a warm-up, and I will turn to AWS for parameter tuning in the next project.

Thanks for all your help!