Black Friday Data Hack

Black Friday Data Hack

AnalyticsVidhya organized a weekend hackathon called Black Friday Data Hack, which was held on 20th-22nd November, 2015.

Black Friday is actually the following weekend, but that's when we've to relax and enjoy :-)

The last hackathon was quite disappointing due to the randomness in the data and the evaluation metric. I was hoping this one would be better.
And it was. Much better.

Problem
The challenge was to predict the purchase amount of various products by users across categories given historic data of purchase amounts.

Data
In general, when you have more data, its always better. The train data had ~ 5.5 lakh observations and the test data had ~ 2.3 lakh observations. The data was very very clean and it feels wonderful to work on such datasets.

The data was of users who purchased products with the amounts. The products had data on three types of categories. The users had data about their age, gender, city, occupation, locality and marital status.

We were to build our models on the train data and score the test data which had pairs of user-product not present in the train data. The evaluation metric was RMSE, which also seemed a very appropriate choice for this problem.

Approach
I spent the first few hours just exploring the data, summarizing variables, plotting graphs, playing around with pivots and in parallel building base models (of course, XGBoost).

On the first day, I was able to go below 2500 with an optimized XGBoost model on raw features. It got me into the Top-3 and since then I've managed to maintain a position in the Top-5.

While checking the variable importance of my XGBoost, I found Product_ID was the most important variable and intuitively it made sense. So, I just submitted the average purchase amount of each product and voila! it scored 2682, which didn't seem like a very bad score. So, all those of you who couldn't cross 2682, here's a simple solution you missed.

Usually ensembles win competitions, but since I couldn't get any model close to the performance of XGB, so I decided to challenge myself to build a single powerful model. Which means, feature engineering.
These two days gave me some wonderful insights on how powerful feature engineering is. With some analysis, gut, trying, cross-validating, here are my final set of features that I used:

Model
User_ID: Used as a raw feature

User_Count: Number of observations of the user

Gender: Converted to binary

Age: Converted to numeric

Marital Status: Used as raw feature

Occupation: Used as raw feature

City Category: One-hot encoded features

Stay In Current City: Converted to numeric

Product Category 1, 2, 3: Used as raw feature

Product_Count: Number of observations of the product

Product_Mean: Average purchase amount of product

User_High: Proportion of times the user purchases products at a higher amount than the average purchase amount of the product

I built an XGBoost with these features, and the code is open-sourced on GitHub, the link is given below.

One very interesting feature I built was
F_Prop: Average purchase amount of product by female users / Average purchase amount of product by male users

This was among the top-3 important variables and gave a CV of ~ 2419 but the LB remained very similar ~ 2430, so I wasn't sure about it. I decided to go without this.

GitHub
View GitHub Repository

Results
This model gave me CV score of ~ 2425 and public LB score of 2428. I was 4th on the public LB, with Jeeban, Nalin and Sudalai in the Top-3. And we finished in the same positions with my final rank being 4th in the private LB.

Views
This is one of the best data-sets I've worked on in a while. The CV and LB scores were perfectly in sync and it was very satisfying to build features and improve the CV as well as LB scores. I'm happy with my performance as I managed to squeeze quite a lot of from the data with a single model.

I might have done better with an ensemble, but just couldn't get anything to work well. And after a while, was just too tired.

Overall, a great weekend, mostly spent on my laptop. For those of you who had memory issues, I worked on my 4GB MacBook Air throughout the weekend. Algorithms and models will advance and become optimized every day, but the power of building good features is still in the hands of Data Scientists like us.
Make the most of it until the machines come and take over ;-)

Thanks to all the folks at AnalyticsVidhya for organizing this hackathon. A big thumbs up from me.

Looking forward to the next Hackathon, and hope it gets better and more competitive.

External Links

View Other Players' Approaches on AnalyticsVidhya

View 3rd place solution code on GitHub by Sudalai Raj Kumar

View 5th place solution code on GitHub by Aayush Agrawal