Problem Set #4

MACS 30000, Dr. Evans
Cheng Yee Lim

Problem 1

Part (a). I created a Kaggle account, username chengyeelim.

Part (b).

Describe a Kaggle Competition of interest to you.

Outbrain is holding the Outbrain Click Prediction Competition to search for algorithms that can better predict recommended content each user will click. With the slew of information available on the web, Outbrain aims to serve as the internets leading content discovery platform and pairs readers with relevant content. Every month, Outbrain has provided 250 billion personalized recommendations from thousands of sites for their users. Thus, the recommendation algorithms are at the heart of their business, which they hope to augment with the brilliant minds on Kaggle.

What is the goal of the competition?

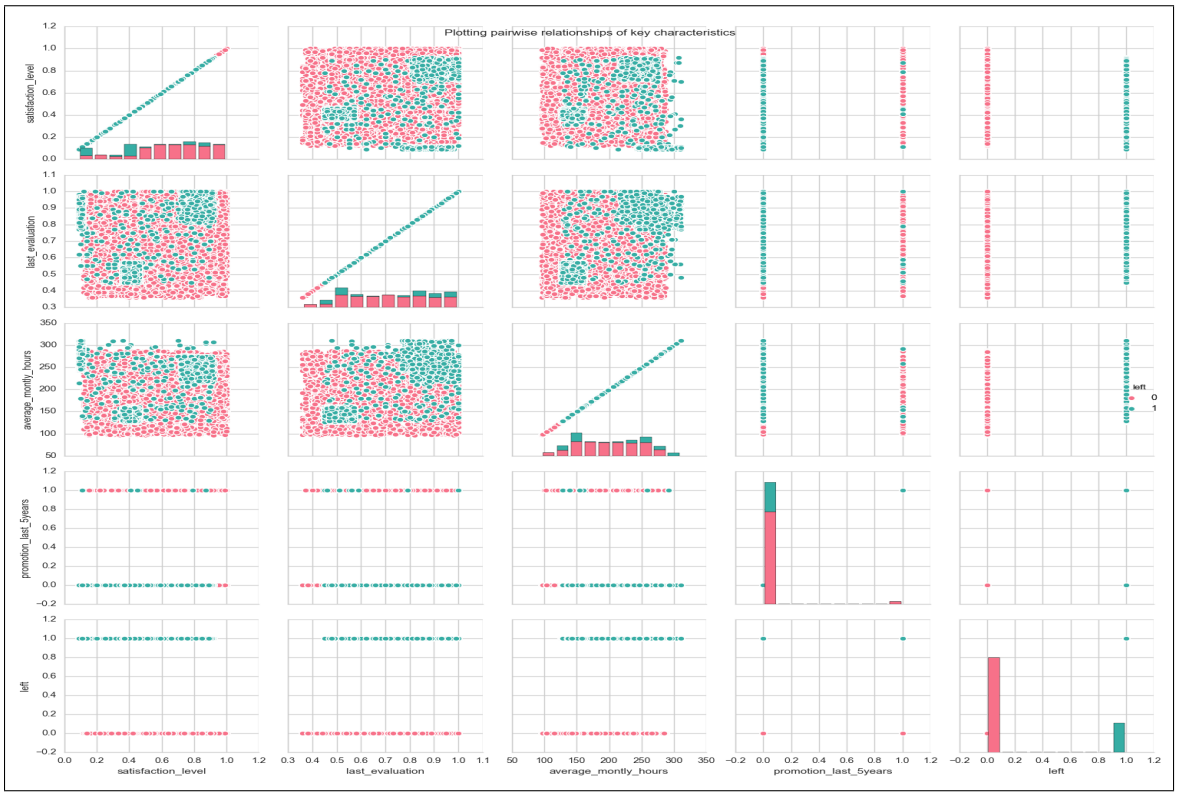
The goal of the competition is to tap on over 13000 active Kagglers to generate better recommendation algorithms that can augment existing their algorithms. An open source call can bring in techniques that have yet to be employed by Outbrains team, or even break through Outbrains group thinking with new perspectives on how to approach this problem. Participants of the competition are graded based on the mean average precision of their code, in predicting whether users will click on certain recommended content.

What would you have to do to make a submission?

To make a submission, I would, firstly, have to accept the official terms and conditions for the competition by January 11, 2017 to comply with all applicable laws and have a binding agreement with Outbrain. Secondly, I have to download the datasets provided by Outbrain to conduct my analysis. Ultimately, I would consolidate my results in a CSV file that entails 6,245,533 predictions and submit it before January 18, 2017. The CSV file will consist a prediction of a list of advertorials ordered by decreasing probability of being clicked for each display_id, which consists of a different number of associated advertorials.

Part (c).

Figure 1: Pairwise plotting of key characteristics affecting employee turnover



Using the Human Resources Analytics dataset from Kaggle, I conducted a preliminary analysis on characteristics that affect employees to leave rather than stay. I began my exploring the data by visualizing the data with a heatmap of the correlations between all the variables, and found that work accidents and number of projects are weakly correlated with leaving the company.

Thus, I did a pairwise plot with the key characteristics, satisfaction levels, average monthly hours, number of projects, last evaluations, promotion over the last 5 years and employee leaving or staying. From Figure 1, we can observe high correlations wherever clusters exist and conclude with the following: those employees with lower satisfaction levels have very high number of hours, Although their last evaluations are high, the employees are not satisfied.

- Employees that left the organisation have lower satisfaction levels, works high number of hours even though their last evaluations were pretty high.
- Most of the employees that left rarely had a promotion over the last five years.

Problem 2

In Montefortes and Morettis recent paper on Real-Time Forecasts of Inflation: The Role of Financial Variables[1] (2013), they attempt to augment inflation forecasting with daily data from financial markets and found that the addition of daily variable reduces forecasting errors, when compared to models that only factor in monthly variables. The issue pointed out by Monteforte and Moretti is that monthly indicators do not capture important information within the month. Thus, they attempt to mitigate the gap by including daily data from financial indicators such as movements in the yield curve or interest rate spread, and combine it with the monthly core inflation index with daily prices of commodities and financial assets.

While their paper is already an improvement in forecasting inflation with more real-time data, the advent of the internet and crowdsourcing platforms enable researchers to collect daily prices of commodities with human computation projects. These days, it is uncommon for individuals to log their groceries expenditure to track their expenses and monitor their dietary intake. Individuals can be encouraged to make it a habit to contribute such information on an open source website. Besides, being a platform to log their groceries expenditure, social elements could be incorporated into the site, where people create sub-threads about tips on eating clean and reducing costs of doing so. Such data can then be collected and replace daily proxies of inflation per month. These newly available data through human computation projects could be an additional input to the Mixed Data Sampling Regression Models (MIDAS), as implemented in Montefortes and Morettis paper, and would be likely to further reduce errors in forecasting inflation.

References

[1] Monteforte, Libero, and Gianluca Moretti. "RealTime Forecasts of Inflation: The Role of Financial Variables." Journal of Forecasting 32, no. 1 (2013): 51-61.