Data Science Project Portfolio

Yumi Kim

1. Industrial Bank of Korea customers' log analysis project. (2019.05 ~ 2019.06)

- **Background:** Increasing usage of smart banking application drives the need to come up with digital marketing plans.
- **Objective:** Cluster the customers who purchased financial products with their log data and come up with marketing plans for each of the groups based on their behavioral patterns.
- **Dataset**: 1) 2.6 million customers' 170 million rows of log data, each of the customers' 2) private information and 3) account transaction data.
- **Project Timeline**

Week 1 ~ Week 4	Week 5	Week 6
Feature Selection/ EDA / Preprocessing	Modeling (Clustering)	Develop Marketing Plans

- **Process**

① Feature Selection / EDA / Preprocessing.
  - Visualized the difference in the behavioral activities between customers who purchased the product and customers who did not.
  - Developed a total of 80 features that captures the behavioral activities before purchasing a product
  - Calculated each of the customers' stay time and number of visits (per page / per time frame)
  - Conducted min-max normalization, one hot encoding, label encoding, sin (cos) encoding according to the characteristics of the features

② Modeling
  - K-Means, K-Means ++, DBSCAN
  - Constructed an elbow point visualization for K-Means model
  - Silhouette evaluation for the result (0.4 accuracy)

③ Marketing Insight
  - Distinguish 5 clusters with their gender and age
  - Derived push alarm message for each of the groups from persona analysis
    Ex) Mid age men tend to visit coupon page 8 to 31 days before purchasing a deposit product. Therefore, recommend mid age men with deposit products who visits coupon page.

2. **Forecast stock return with ARIMA / regression analysis (2018.11 ~ 2018.12)**

- **Github Code / Presentation Slides**
- **Background:** Find the factors that impacts the stock price and make investment decision with portfolio optimization
- **Objective:** Forecast 2019 stock return for companies that comprise Dow Jones Index
- **Dataset:** 41 companies' monthly stock return, google trend data, dollar index, PMI, Federal Reserve, and unemployment rate from 2004 Jan to 2018 Nov.
- **Process**

Step 1	Collect data
Step 2	Set investment strategy: · Initial investment is $100,000, and profits from the backtesting period are reinvested. · Forecast 1 year forward return every year on December 1st and select top 5 companies with highest expected return. Optimize portfolio on the maximum sharpe ratio point.
Step 3	(Data from 2004 Jan – 2013 Dec) ARIMA modeling: Forecast 1-year forward values for the 5 factors Regression modeling: Predict companies' 1-year forward return by regression analysis with the factors predicted from ARIMA modeling.
Step 4	(Data from 2013 Dec – 2018 Nov) Backtesting: Repeat step 3 for each of the year
Step 5	Investment decision for 2019

- **Evaluation** (4-years period of backtesting)

① **Annual return** (based on initial investment)

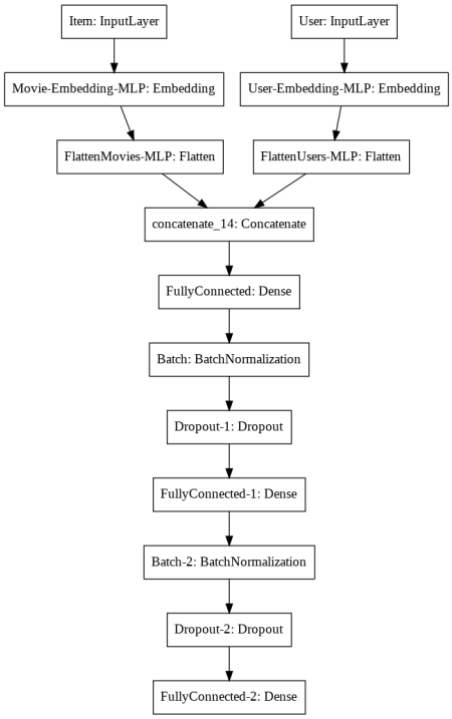
2014	2015	2016	2017	2018
5.8%	17.1%	21.3%	39.2%	12.6%

② **Detail**

	2013-12-01	2014-12-01	2015-12-01	2016-12-01	2017-12-01	2018-12-01
Current Portfolio (weight %)	C (47%) BAC (43%) AIG (5%) AA (0%) CAT (0%)	BAC (53%) AIG (37%) TRV (0%) C (0%) AA (0%)	AIG (74%) C (19%) HON (0%) AA (0%) TRV (0%)	AIG (60%) HON (17%) AA (12%) TRV (0%) C (0%)	AIG (42%) AA (35%) HON (14%) TRV (0%) C (0%)	X (Sell everything)
Gain	X (initial investment)	$5,834	$11,308	$4,110	$ 17,956	X
Loss	X (initial investment)	X	X	X	X	$26,567

3. **Recommendation engine for movies (2019.08 ~2019.09)**

- **Github Code**
- **Objective:** Predict the ratings of the movies for each of the users with deep learning recommendation system.
- **Dataset:** Movielens
- **Process:** Applied model from the Neural Collaborative Filtering paper and made latent vectors for users and movies and concatenated the vectors.



4. **Plant image classification (2019.03)**

- **Github Code**
- **Objective:** Predict 12 different classes with the plant image.
- **Dataset:** Kaggle Plant Seedings Classification
- **Process:** Processed the image by deleting redundant background and conducted CNN modeling with keras library.