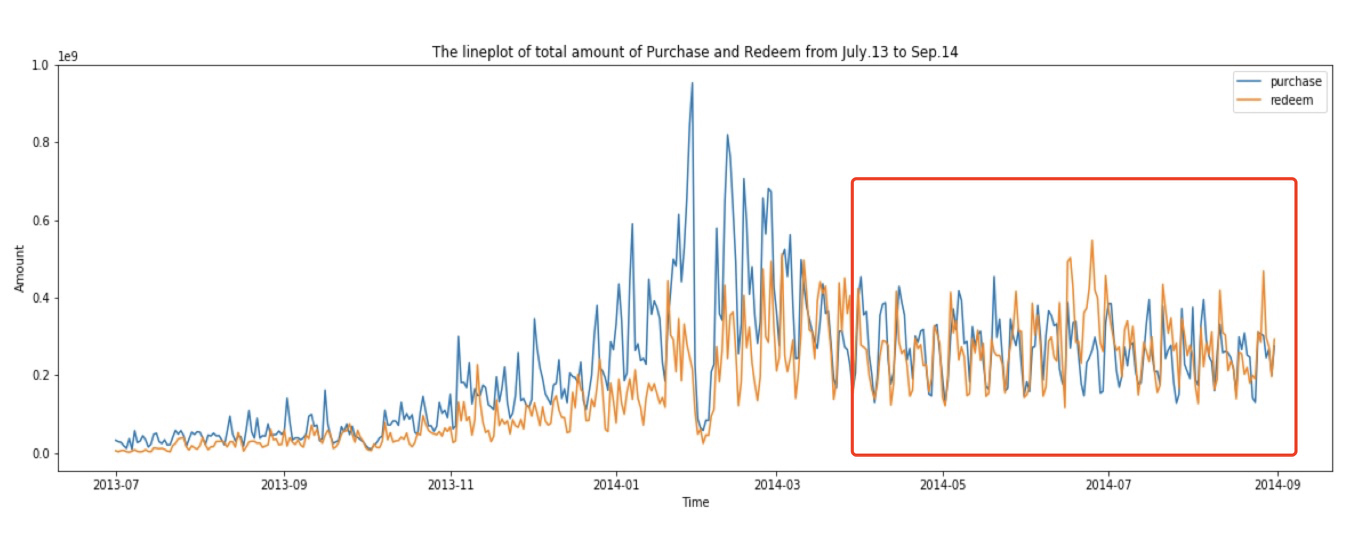
4. Feature Engineering

A time series chart of the total daily purchases and redemptions every day from July 13 in 2013 to September 14 in 2014 was drawn after the sum of 100,000 users. The results are shown in the figure below.

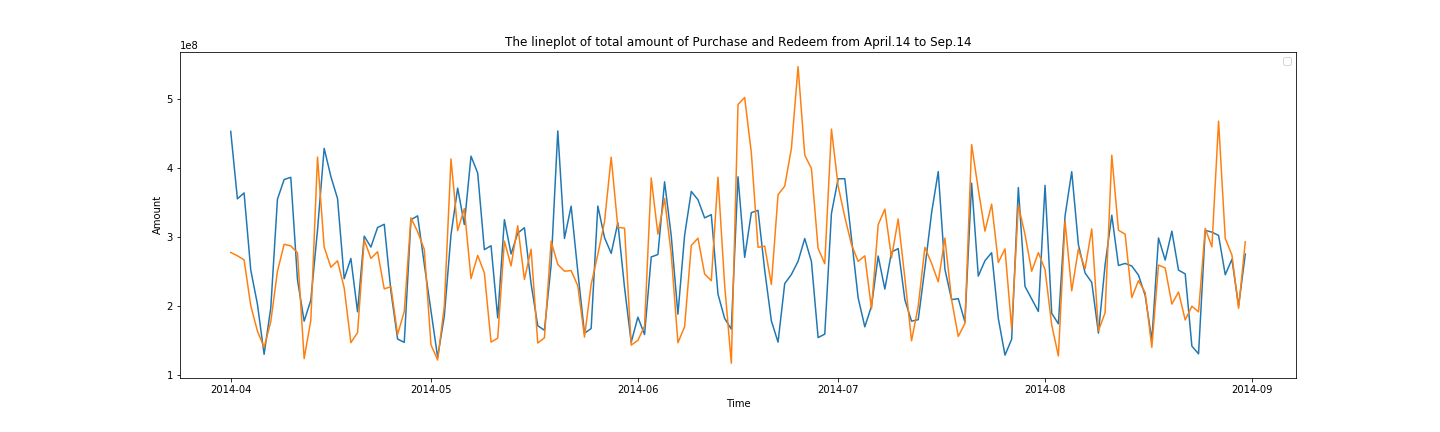


By observing the figure above, we can find that the overall data have significant characteristics:

1. Obvious periodicity on a seven-day basis.

2. The amount of working days is relatively high, while the amount of holidays is relatively low.

3. The overall trend is relatively low from July 2013 to Nov. 2013, relatively large fluctuations from Nov. 2013 to Apr. 2014, and a stable period after Apr. 2014, which lasts until the end of Aug.2014.

To model the data for September, we need regular user behaviors requiring data to be relatively smooth. In addition, we need to select samples which can describe the characteristics of September. 

Combining the above cyclical and holiday characteristics, we can construct a static feature of the date.

For time series, we will construct “is” feature based on date static feature and “distance” feature separately.

As for macroeconomic factors, since monetary funds are main participator on the inter-bank market, so we will consider Shibor and Balance of Margin Trading and Securities Lending as features.

5. Feature Selection

We divided feature selection process into 5 steps: Low separation ability, Multi-collinearity, Low correlation, Shapley value and Permutation Importance.

Feature selection is the process of selecting a subset of relevant, useful features to use in building an analytical model.

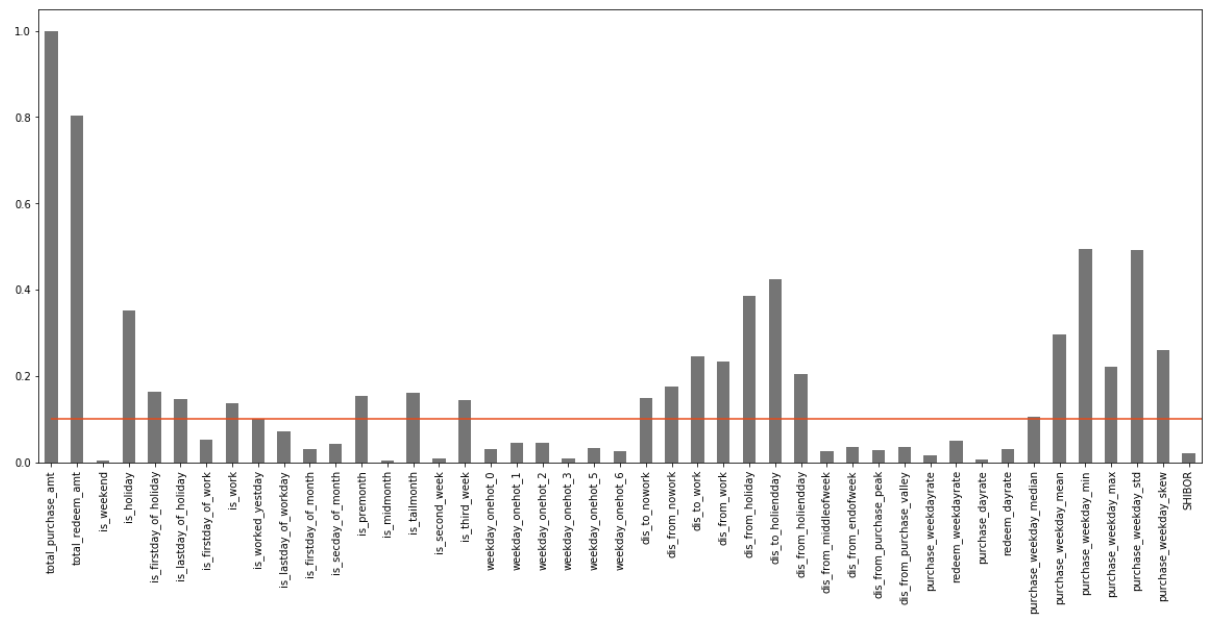
Firstly, we delete features with low separation ability.

Secondly, we delete features with multi-collinearity. Rules :

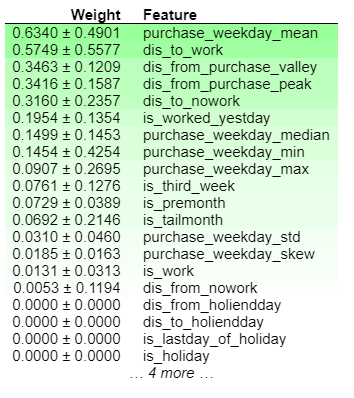
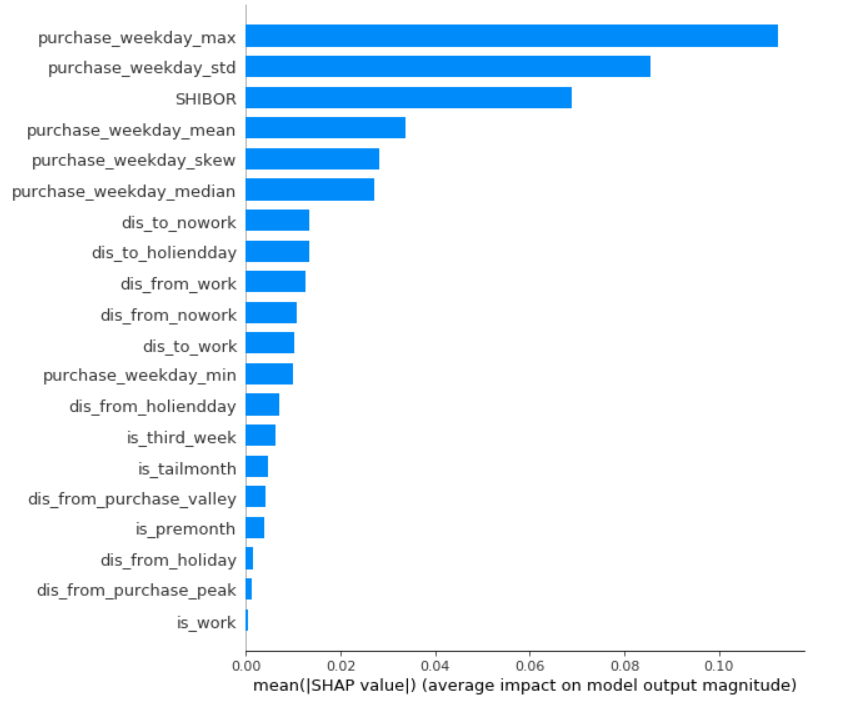
Compute correlation coefficient between any two features. If two feature’s correlation coefficient is larger than 0.8, then delete the one with small correlation with predicting variables.

Use VIF statistic value to detect potential multi-collinearity, the truncated value is set to be 10.

Thirdly, we delete features with low correlation. We set the threshold as 0.1.



Fourthly, we select the features with high Shapley value. Fifthly, we select the features with high permutation importance.



Sixthly, we get the intersection of these two selection method.

References

[1] Belsley, D. A. (1991). Conditioning diagnostics: Collinearity and weak data in regression (No. 519.536 B452). New York: Wiley.

[2] Chatterjee, S., & Price, B. (1977). Selection of variables in a regression equation. Regression Analysis by Example, 201-203.

[3] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpret- ing Model Predictions. In Advances in Neural Information Processing Sys- tems 30. Curran Associates, Inc., 4768–4777. http://papers.nips.cc/paper/ 7062- a- unified- approach- to- interpreting- model- predictions.pdf

[4] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang New- man, Jerry Kim, et al. 2017. Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. bioRxiv (2017), 206540.

[5] Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

[6] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

[7] Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 26(10), 1340-1347.

[8] Achard,S. et al. (2005) Criteria based on mutual information minimization for blind source separation in post nonlinear mixtures. Signal Processing, 85, 965–974.

[9] Strobl,C. et al. (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8, 25.

[10] Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. J. Mach. Learn. Res., 3, 1157–1182.

[11] Breiman,L. et al. (1984) Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

[12] Breiman,L. (2001) Random Forests. Mach. Learn., 45, 5–32.

[13] Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. J. Mach. Learn. Res., 3, 1157–1182.