Time Series Anomaly Detection for Credit Card Delinquency

Data Science Institute
COLUMBIA UNIVERSITY

Authors: Feihong Liu, Han Ding, Yunbai Zhang, Zichen Pan, Zilin Zhu
Industry Mentors: Gerry Song
Faculty Mentors: Adam S. Kelleher

Data Science Capstone Project with Capital One

Introduction

Nowadays, a lot of people cannot afford their credit card debt or even minimum payment. It will be a big loss for banks if they cannot detect credit card delinquency in advance. We hope to get time series anomaly detection algorithms to help bank do this thing according to customers historical transaction data. There are many algorithms to do time series detection, but little meta-analysis has been done to compare different approaches. Therefore, we have trained several traditional machine learning algorithms and deep learning algorithms on manually generated data and compared their performances.

System Design

The main design of our system consists of 2 parts: data generation and algorithm evaluation. Due to credential problems, we won't have access to real data. We proposed a statistical model for credit delinquency.

- **Our generator is based on Hidden Markov Model with modifications.**
- **Hidden variable(3): Normal, Struggling, Abnormal. (the financial state of a customer)**
- **Observation(3): a customer behaviors by the end of a month, paying all the debt, only the minimum payment (to avoid the interest) or nothing**
- **Goal: detect when a customer will turn abnormal.**
- **Modification: to vary the character of users, we generate different transition and observation matrices for different users using a priori distribution.**

To do the algorithm evaluation, we cut the data into fix length vector pairs (both hidden variable and observation) and label the vectors whose hidden variable will turn to abnormal within some future timestamps as positive and other as negative. Then, we will remove the hidden variable vectors and evaluate the algorithms using the observation vectors and label.

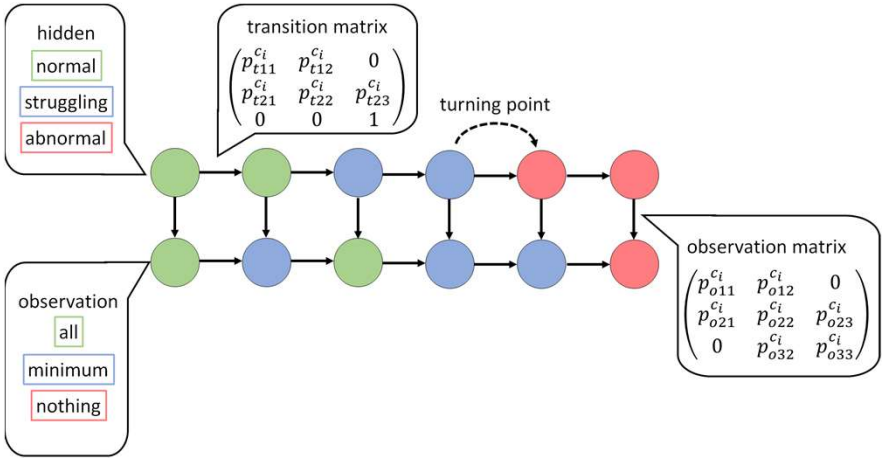

Figure 1. Data Generating

Result

Based on Precision-Recall curve of classic machine learning models, logistic regression, SVM and XGBoost perform better when recall is lower than 0.6. After recall reaches 0.6, all the models perform similarly.
Based on Precision-Recall curve of NN-based models, LSTM and LSTM with attention performs better than others when recall is lower than 0.3. After recall reaches 0.3, all of the neural network models perform similarly.

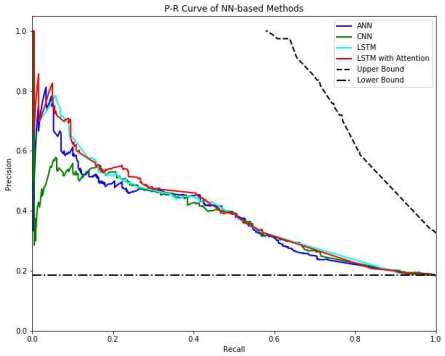

Figure 2. P-R Curve of NN-based Methods

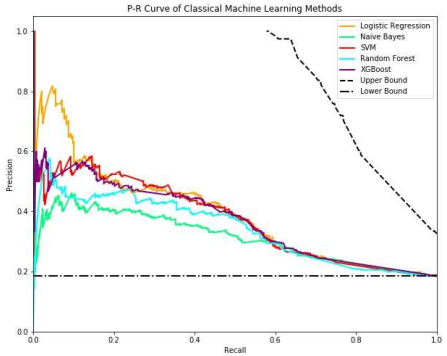

Figure 3. P-R Curve of Classical Machine Learning Methods

Conclusion

For the future work, we plan to improve our data generation process to be more specific, such that it would fit policies of different banks.

Acknowledgement

We thanks for the support , ideas and effort of Professor Adam S. Kelleher and mentor Ms. Gerry Song.

Reference

[1] Yahoo! Webscope dataset ydata-labeled-time-series-anomalies-v1_0. http://labs.yahoo.com/Academic_Relations
[2] The Numenta Anomaly Benchmark (NAB). https://github.com/numenta/NAB