Q&A Feature Extracting Framework for Online-Lending Collection Risk Modeling with X-Encoder*

Abstract

SongTao Jiang
CreditX
ShangHai, China
jiangsongtao@creditx.com

Wei Min
CreditX
ShangHai, China
minw@creditx.com

Qiang Gao
CreditX
ShangHai, China
gq@creditx.com

ABSTRACT

Online lending market in China has grown rapidly since 2016. According to reports, the leading financial institutions are receiving millions of applications and granting loans for hundreds of thousands customers each day, creating over 100 millions yuan transactions. However, growth of the market leads to the growth of credit risk. In addition to increasing efforts on application fraud detection, financial institutions are facing much more collection pressure as well. Aiming to better estimating the quantified post-loan collection risk, we propose a new Q&A based feature extracting framework, with Q&A texts generating from audio collection call records, to identify the willingness of a debtor to repay his debt. Such framework is called X-Encoder based on deep neural networks. The results on massive data show that such framework can improve the sufficiency of modeling collection risk significantly, increasing the capability in quantifying collection risk at least 50% compared to traditional models on texts.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**;

KEYWORDS

online lending, financial fraud detection, post-loan collection, deep learning, feature extraction, X-Encoder

ACM Reference Format:
SongTao Jiang, Wei Min, and Qiang Gao. 2018. Q&A Feature Extracting Framework for Online-Lending Collection Risk Modeling with X-Encoder: Abstract. In *Proceedings of WSDM workshop on Misinformation and Misbehavior Mining on the Web (MIS2).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Online lending in China, comparing with payday loan in America, is a small, short-term unsecured loan, regardless of whether repayments of loans are linked to borrowers' paydays, providing financial support to less-qualified customers. On the other hands, these less-qualified customers has no guarantee nor mortgage, and uncovered by traditional credit service, causing the overdue ratio of payday loan estimated over 20% and considerable related loss for the financial institutions. Hence, establishing proper strategies during collection process is taken into more consideration for these institutions. Specifically, if a collector identify the debtors' personal characteristics well, precise description for the debtor can be established and targeting collection decisions can be easily made. Although the techniques of risk modeling developed rapidly, decision-maker are still facing several obstacles [1].

Regulation for online personal privacy According to the newly came-up regulation in China, neither the website nor the mobile application is allowed to collect personal information except for necessary identification for individuals. As a consequences, for online financial institutions, data shortage is the primary problems that collection risk models are facing[2].

Market risk & information asymmetric The growing market provides considerably many choices to debtors, causing an abuse in loaning applications. While there are limited ways to collect data from a debtor, financial institutions are not able to sufficiently rate the risk level for the loan. In many default cases, up to 40% of total defaults, debtors are more likely to borrow from another lending company to clear the current debt yet this can be concealed with a minimum effort due to the insufficient market. Loans related with such repayment methods reflect lower financial stabilities and higher risk comparing to loans with normal repayment methods.

Lack of measurements on repayment willings One of the severest situations in loan collection process is that most of the default cases are not caused by lack of money but debtors' unwilling to repay. Most of the default cases cannot be reported to Chinese Central Credit System so that no penalty on fraudster is performed, regarding of a low cost on defaults.

In Chapter 2, we will introduce our framework, X-Encoder; in chapter 3, we will describe how we applied X-Encoder practically and exhibit the results of experiments.

2 PREVIOUS RESEARCHES AND METHODOLOGIES

2.1 Problems Definition

To improve the data sufficiency of collection risk modeling, we are aiming at collecting the audio records of collection calls on specific debts. With converting audio data into text data, the problems of identifying willingness of repayment and potential fraud can be turned into speech recognition problems. Also, researchers among the industry have already done tremendous empirical studies on

*Produces the permission block, and copyright information

MIS2, 2018, Marina Del Rey, CA, USA

ACM ISBN 123-4567-24-567/08/06.
https://doi.org/10.475/123_4

the pattern that a phone calls might represent within a collecting conversation.

Furthermore, what should be distinguished from traditional speech recognition is that the phone call texts consist of interactive conversations. Not only speech recognition of debtors' speech but also collectors' speech is valuable. The logic and speech representation are dynamic for both parties while the conversation is going on. Thus, the dynamic interaction for both parties should also be taken into consideration when constructing risk models[3]. There are several patterns found by previous scholars that impact the speech recognition during the conversations.

Threatens from the collectors Threats from collector tend to push debtor to repay. Collectors threat the debtors that the nonpayment of any debt may be resulting in the arrest or imprisonment of any person or the seizure, garnishment, attachment, or sale of any property or wages of any person unless such action.

Attempts from the debtors Most of the default cases, debtors are willing to repay the loan. It is crucial that we separate these debtors from the fraudsters.

Unreachable debtors & Vicious premeditated default Significant cases shows that whether the debtors are viciously acting defaults can be identified from their tones and phrase using.

2.2 Fundamental Methodologies

2.2.1 **Word to Vector**. Word to vector method will be the foundation of our feature extracting framework, X-Encoder. Several text representations have been used during past few decades such as simply converting bag of words into sparse matrix. However, in the real cases, the matrix converted from bag of words can have millions of dimensions, causing serious dimensionality problems which reduces the efficiency of model training and model performances. Thus, based on the logic of dimensionality reduction, word to vector plays an important role in speech recognition especially in speech recognition with deep learning and was highly recommended by previous researchers. Furthermore, in practice, one may want to introduce some basic pre-processing, such as word-stemming or dealing with upper and lower case. In previous experiments, word2vec can capture text representations very well [6][7][5].

2.2.2 **Auto-encoder**. An auto-encoder is a neural network that is trained to attempt to copy its input to its output. Internally, it has a hidden layer **h** describes a code used to represent the input.Additionally, an auto-encoder succeeds in simply learning to copy the input everywhere, then it is not especially useful. Instead, auto-encoders are designed to be unable to learn to copy perfectly to increase its validity[5].

An auto-encoder always consists of two parts, the encoder and the decoder, which can be defined as transitions ϕ and ψ such that:

$\phi : \mathcal{X} \rightarrow \mathcal{F}$

$\psi : \mathcal{F} \rightarrow \mathcal{X}$

$\phi, \psi = \arg\min_{\phi,\psi} \|X - (\psi \circ \phi)X\|^2$

2.3 X-Encoder

As discussed above, phone calls can be more complicated than a review or an article since the representations for both parties vary dynamically. Thus, the speech recognition should be examined conditionally. In X-Encoder, the representations of collectors' texts

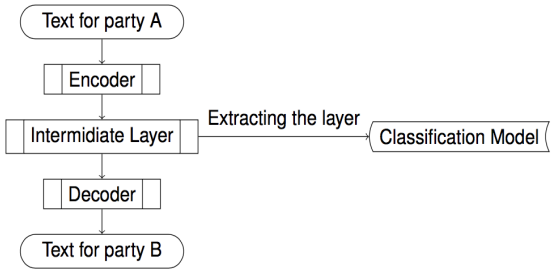

Figure 1: auto-encoder framework

are processed along with the presentations of debtors' texts and vice versa. While the classic auto-encoder is seeking the inner transition within a corpus, X-Encoder recognizes both parties separately. Here we found two examples showing distinctive presentation with same words [1]:

Senario 1

Collector : Would you be able to repay the loan within 3 days?

Debtor : I'd repay it this afternoon if that's ok. → Positive repayment willingness.

Senario 2

Collector : Your loan has been overdue for 20 days. You have to repay it now.

Debtor : I'd repay it this afternoon if that's ok. → Negative repayment willingness. [2]

X-Encoder is generated based on the logic auto-encoder discussed in 2.2.2. Massive previous researches aimed to evaluate or predict the quality of questions and to find high quality answers[8] [9] . However in X-Encoder, while we reduce the dimensionality of text representation, we are not seeking for a decoder to copy the original text completely. Instead, we use the text representation for another party as output to find the presentation that can describe the relationship between the input texts and the output texts. The structures of X-Encoder is shown in figure 1.

The input of X-Encoder will be collectors' texts or debtors' texts and the output will be texts for another party. Compared to traditional NLP framework, word representation recognized by X-Encoder is not an isolate interpretation of its distribution in the whole corpus. In fact, on the contrary, X-Encoder adjusts the recognized sentiment by taking different scenario into consider.

3 EXPERIEMENTS

3.1 Sample Definition & Data

Our dataset consists of 50000 samples with 25000 positive samples and 25000 negative samples. The overdue days [3] in these samples vary from 1 day to 20 days distributed evenly. We define our analysis target as whether this loan will be repaid in ten days. If a client repays within 10 days start from the time he is recorded, he is marked as a good client (target = 0), otherwise he is marked as a

[1] Originally in Chinese, here we translated into English

[2] based on collectors' footnotes

[3] overdue days means how many days have past since the day that this loan should be repaid

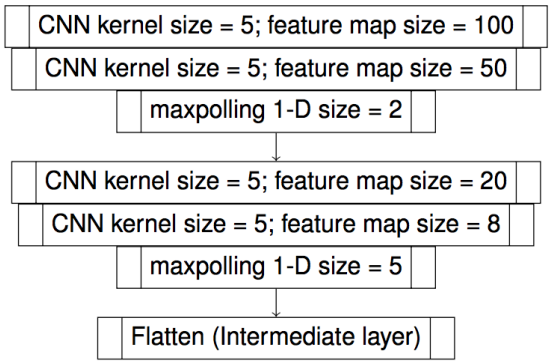

Figure 2: Encoding framework

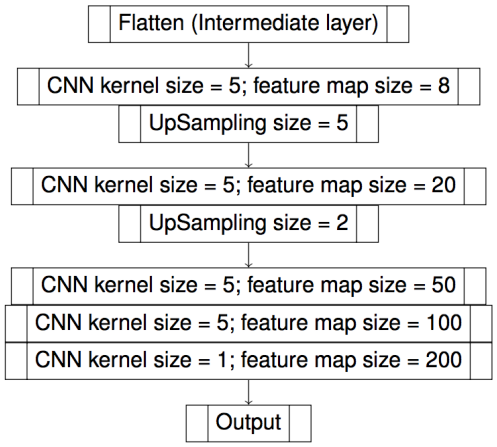

Figure 3: Decoding framework

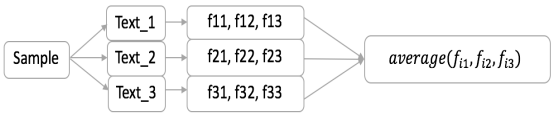

Figure 4: Feature Integration

bad client (target = 1)[4]. The timestamp when a samples generated is defined as **collection date**. Sample client might occur several times with different collection date. We also collected all the texts for collection call records, converted from audio records. There are **235,137 records** in total and 302,000 of them can be matched to a certain client after filtering calling date and collection date.

3.2 Feature Extracting with X-Encoder

To extract features with X-Encoder, we firstly define the format of our texts.

Collectors' texts and Debtors' texts word2vec size = 200; filtering 50 words; for each text the output vector will be 50 x 200

To choose 50 words from original text, we use tf-idf to filter the words that occurs too often Then we define the structure of our encoder and decoder. We shared the same encoding-decoding structure in X-Encoder.

The structure of these framework will be as followed.

Encoder *Figure 2*

Decoder *Figure 3*

We trained our X-Encoder with rmsprop. Since we've already vectorized the texts, it's better to choose linear activation to obtain negative value during the training process. Our loss function will be mean squared error function. Two X-Encoders will be employed. One with collectors' texts as input data and the other one with debtors' data. Each X-Encoder will generate 40 features (40 neuros from the intermediate layers) . The last step of feature extracting with X-Encoder is to integrate feature from text into an individual sample, as figure 4 shows.

3.3 Compared Feature Extracting Methods

To make a comparison between X-Encoder and traditional modeling methods, we construct several other models.

Feature generated manually

Speech recognition can be pre-defined by humans using out own expertise. Here we extract several features by statistical methods and regularizing text. e.g:

Average_void_duration = total void time (in seconds) divided by total calling counts. Capturing the information that debtor is not able to explain or responding to collectors questions.

Average_cursing_count = total sentences counts containing curing words / total calling counts. Capturing whether either party uses cursing words.

After manual extracting features from original text, we throw them into Gradient Boosting Machine by splitting 20% samples out of total samples as test set. In advance, we spare samples after August 15th as out of date samples.

Feature generated by LDA models

Latent Dirichlet allocation, known as LDA model, is a generative statistical model widely used in nlp for speech recognition. The target of LDA model is to cluster all the text and the results of LDA will be used in classification model[4]. As a popular speech recognition model, LDA has been proven effective by previous scholars. Thus, we trained several topic models to cluster all the collection text, model with 10 topics, model with 20 topics and model with 50 topics. Also we integrate features using the distribution of a collection text on different topics as feature for this specific text, and then averaging the probability over the same topic among all the text that matched to a certain client sample. Therefore we got 10,20 or 50 features for a client sample representing 'his' probability while responding collectors on different topic.

[4]The samples are not defined in the end of a collection period. As a matter of fact, same client can appear in the sample set duplicately. This is due to the dynamic state of a client's credit status and multiple collection strategies. For example, if one repays after overdue 20 days and was called for collection at 5th days and 15th days, then two circumstances should be treated differently, bad client and good client respectively

Table 1: Feature Importances

Feature Ranking	Manual iv	LDA iv	DL Framework iv
Top 1	0.1170	0.0305	0.4462
Top 2	0.1104	0.0277	0.4384
Top 3	0.1104	0.0103	0.4007
Top 4	0.1056	0.0090	0.3784
Top 5	0.0875	0.0056	0.3738

Table 2: Model Results

Model Feature	Train set	Test set	Held-out set
LDA	0.556 (0.069)	0.542 (0.062)	0.551 (0.084)
manual + LDA	0.642 (0.202)	0.632 (0.196)	0.642 (0.201)
X-Encoder	0.724 (0.326)	0.701 (0.296)	0.702 (0.311)

We put them into Gradient Boosting Machine as well. In this process we keep three part of the data, train, test and held-out test same as how we spit in manual feature model.

3.4 Results

3.4.1 Feature Importance (information value). The abilities of features to classify the samples are prerequisite to the modeling performances. Here we present the information values for three parts of features, manual features, topic features and out deep learning framework features.

As we expect, feature importance results (Table 1) show that features from our framework outperform twice as much as others measured by information value. However, one of the interesting findings is that features generated from topic model didn't outperform manual features by human expertise. Carefully going through the data, we found that the text converted from call records are not as clean as an normal article such as a report from a newspaper. The process of conversion is depended on the accent of debtors (while the collectors are ordered to speak madarian), the quality of the call records (influence by the signal for example) or the accuracy of the conversion. LDA model cannot handle such noises as expected, while both manual and deep learning framework can resolve these noises in a better way. For manual feature, we did not care about the whole text as much as LDA models, only whether there are occurrences of certain words or the properties of the call itself (e.g. call duration). On the other hand, word to vector reduce the dimensionality of the texts into a certain vector spaces, which is the reason why it's stable than LDA using word bags.

3.4.2 Model Performance. Table 2 gives us the results of model performance [5], it is not hard to see that out framework increase max-ks, an industry standard metric measuring model performance, has increased nearly 50% from 0.196 to 0.296.

Also, test results on out-of-date test shows the stability of all three models, thus the comparison among these models are validate with respect to the presence of over-estimation. Here we also check both model PSI and feature PSI to evaluate the stability of our framework. With our CNN based Q&A nlp framework, feature PSI between out-of-date samples and training samples are all below 0.1, comparing to 0.35 on average for manual features.

[5]The results presenting in the table consist of two parts, roc and max-ks in the bracket. For example, 0.556 (0.069) shows that the model has 0.556 auc and 0.069 as max-ks

4 CONCLUSION

In this paper, we presented a validate and sophisticated method to increase the ability of feature extraction in post-loan risk modeling. We started by explicating the obstacles that Chinese online financial institutions are facing and situation of data shortage when we construct post-loan risk models. But with using call records between collectors and debtors legally, by applying our X-Encoder, financial institutions can reduce their misjudgement on willingness of a debtors and pay more attention on collecting resources allocation to reduce the losses. In sum, the X-Encoder we introduce can be an effective way to detect fraud when online financial institutions collecting their overdue loans.

REFERENCES

[1] Insley Jill. *GE Money refuses mortgages to payday loan borrowers.* The Guardian. London (2012/07/12).

[2] CHINA *Section VII (2), Provisions on the Administration of Mobile Internet Applications Information Services* (2017).

[3] Daniel Tam. *How To Spot A Payday Loan Collection Scam.* Avvo (2012/07/03).

[4] David M.Blei, Andrew Y.Ng and Michael I.Jordan. *Latent Dirichlet Allocation* (2003).

[5] Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning* MIT Press, (2016)

[6] Ronan Collobert and Jason Weston. *A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning* (2008).

[7] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. *Learning Word Vectors for Sentiment Analysis* Stanford University.

[8] C. Shah and J. Pomerantz. *Evaluating and predicting answer quality in community qa.* In Proc. SIGIR, pages 411?418, 2010

[9] Zhang J, Kong X, Luo RJ, Chang Y, Yu PS. *Ncr: A scalable network-based approach to co-ranking in question-and-answer sites.* InProceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management 2014 Nov 3 (pp. 709-718). ACM.