Feature Generation Using Machine Learning from Large Sparse Financial Data

*Emergent Research Forum (ERF)*

# 1. Introduction

[In credit card business, marketing offering is a key analytical focus.

not only … but also credit transactions such as loan eligibility, interest rates, bad debt prediction, personal finance management,

customized marketing, tailored,

analytical challenge is that

improve target accuracy such as recommended coupons’ redemption rate ….. targets and non-targets ?discover new targets ?

computational modeling of customer features is a core component of financial data analytics.

Hence, feature processing has become important]

[predictor variable….Non-normality …… Measurement bias

skewed …. where non-zero values are scarce

very sparse binary variable from a rich data of

Consequently, makes prediction unreliable? A second characteristic of data in a credit card domain is that whereas payment records are reliable, customer information such as demographics and …. behavioral …

rich in volume but not rich in dimensions……. how to determine which is noise and which is not.

Commensurate, fractions of the original value

In contrast to online businesses where factors that affect purchase such as willingness-to-pay should be estimated from proxies outside of the focal domain , exogenous to the model]

[to predict the target *y*,dependent variable…. Another caveat in financial data is Variables with low levels. that classification targets are scarce. For example,

In binary classification, traditional methods such as logistic regression is not suitable for parameter estimation. classification problem, regression problem

induction algorithms that scale well to find, among a large amount of correlated variables, comparably irrelevant ones to be eliminated to better predict the market indices.

Problems such as perfect separation may occur.

#### class imbalance is a classic problem in finance. True classes are scarce. economic events in market, insurance claims from clients, frauds among normal transactions, and defaults in banking, to name a few. Our targets constitute only 2% of the population. A highly imbalanced class may hinder adequate fitting in parameter space, so adjustments are also made in the experiment.

parameters for classification,]

[In this study, we test feature preprocessing using multiple machine learning approaches in combination with established methods as part of a prediction pipeline. ]

# 2. Related Work

Characteristics of financial data and machine learning operability

does not scale well

bootstrap aggregation (bagging) classifiers and boosting classifiers.

There is still yet scare research on learning problems in financial applications. López de Prado

empirical

such as support vector machines (SVM), clustering, matrix factorization

The statistics literature …. penalized regression such as Lasso (Tibshirani 1996), Elastic Net (Zou and Hastie 2005), elastic net which combines L1-regularization and L2-constraints

L1-penlaty linear Support Vector classifier

These include neural networks whose internal nodes are feature extractors. neural networks such as Restricted Boltzmann Machines (RBMs)

In Bernoulli restricted Boltzmann machines, all units are binary stochastic units, meaning that the input data should either be binary or between 0 and 1.

Despite the large volume,

due to

multinomial labels

binomial labels and numerical attributes

begets a problem of sparse data

Due to the scarcity of financial transaction data,

Other applications include management science (Ban and Rudin 2018).

While in data mining …

Supervised unsupervised

In this paper we use unsupervised training with Restricted Boltzmann Machines (#######) to transform the feature space. Recently, several algorithms have been developed, which are able to learn features from datasets in different domains [5, 6, 10, 9, 17, 25]. In applications to computer vision, the state-of-the-art feature learning showed similar or better performance compared to non-learning algorithms [10, 9, 17].

# 3. Methods and Models

## Feature Space Transformation

A restricted Boltzmann machine is a two-layer connectionist system representing a joint distribution *P*(*v, h*) of states of units in visible layer *V* and hidden layer *H*. Normally, for learning feature representation, the observed data is encoded in visible layer and the outputs in hidden layer are considered as learned features.

advantage of using RBMs….

composite feature space

dimensionality reduction, principal component analysis (PCA)

## Resolving Class Imbalance

random undersampling, random oversampling, SMOTE, TOMEK,

Sample weights,

One potential method is sampling: subsampling, oversampling of the minor classs, undersampling of the major class,

## Model Training

ensemble methods, a collection of weak learners

For extracting features, support vector machines

random forests, extra trees, rbm,

for validations, logistic regressions, random forests, extra trees, bagging, adaptive boosting

through voting

# 4. Experiments

## Data

The experiment uses customer profile and consumption records from one of the major credit card companies in Korea. Transactions from August 2018 through October 2018 were aggregated to the card holder level. The data set was limited to 100,000 randomly chosen individuals who have at least one credit card usage in selected businesses during the window. First, about 644 measurements were chosen based on heuristics. The initial list of variables include customer’s age, sex, occupation codes, address area codes, internally managed customer grades, etc., most of which are multi-class variables. A few more binary variables were introduced, for instance, indicating whether the customer holds a specific kind of card or not. Additionally, count variables were considered, e.g., the occurrence of specific types of transactions. The remaining involves numerical metrics derived from credit line and amount used, divided into whether it was spent on weekdays or weekends, per merchant category. The encompassing categories are thus *N* (number of merchant business types) × *M* (number of months) × 2 (weekends or weekdays). Finally, behavioral proxies to capture were added.

Then, categorical variables were transformed through one-hot encoding and continuous variables were scaled using standard normalization (i.e., *z*-score, see below). Missing values were either encoded as a separate class or substituted by column mean.

After preprocessing, a total of 774 features were produced. Cross-correlation was checked and no pair showed strong collinearity or interdependency. The total observation is therefore 77.4 million, which majorly contain zeros.

## Feature Extraction

The experiment evaluated how the performance of learning models is affected by using the RBM feature transformation.

In this experiment we evaluate how the performance of similarity learning models is affected by using the RBM feature transformation as a preprocessing step. First, the parameters of the RBM and its unsupervised training are explored and then fixed to compare results with the data published in [28] using the same 10-fold cross-validation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Linear Model | | | Tree-based Model | |
|  | *L*1-SVM | *L*1-SVM (balanced) | *L*1-Logistic | Random Forest | E |
| Total Number of Features Selected | 64 | 178 | 230 | 232 | 329 |
| Hyperparameters | alpha | alpha | C = |  |  |
| Table 1. Comparison of Feature Selection Models | | | | | |

The visualization in Figure 1 confirms the two approaches differentiate survived features.

[Insert Figure 1 about here ]

[Insert Figure 1 about here ]

[Insert Figure 1 about here ]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  | LSVM | LSVM\_bal | Logistic | RF | XT |
| *L*1-SVM |  |  |  |  |  |
| *L*1-SVM (balanced) |  |  |  |  |  |
| *L*1-Logistic |  |  |  |  |  |
| RF |  |  |  |  |  |
| Extra Trees |  |  |  |  |  |
| Table 2. Cross-correlations of Feature Importance | | | | | |

Table 3 shows the measure of similarity among the different selection results. Again,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  | LSVM | LSVM \_bal | Logistic | RF | XT |
| *L*1-SVM |  |  |  |  |  |
| *L*1-SVM (balanced) |  |  |  |  |  |
| *L*1-Logistic |  |  |  |  |  |
| Random Forest |  |  |  |  |  |
| Extra Trees |  |  |  |  |  |
|  |  |  |  |  |  |
| Table 2. Cross-correlations of Feature Selection | | | | | |

## Evaluation

Evaluate the performance of the predictor as part of a pipeline, 50% of data set was set aside for validation,

individual predicted probability of acceptance (*score*) was calculated based on

a priori

evaluation metrics

Table 4 summarizes the predictive performance of the selected models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| Logistic |  |  |  |  |  |
| SVM |  |  |  |  |  |
| RBM + Logistic |  |  |  |  |  |
| RU + Bagging |  |  |  |  |  |
| RU + RF |  |  |  |  |  |
| Table 4. Predictive Performance of Chosen Predictors | | | | | |

# 5. Discussion

[Effects of fitting methods on feature selection models]

[Effects of regularization on feature selection models]

[Effects of sample weights on feature selection models]

[Effects of scaling on feature selection models]

[Effects of dimensionality reduction on feature selection models]

Bagging parallel, boosting sequential

Regularized …. tree-based approaches perform present more consistent results

The contrast between linear and tree-based model does not only demonstrate the difference in the degree of conservativeness (to which the variables are shrunk to zero) but also illustrate the potential impact of mathematical properties of variables on selection results.

gradient vanishing problem, in which optimization does not converge in parameter space

hybrid of

Predictive model as part of a pipeline

a priori

highly localized, not versatile

One can also note that many methods of variable subset selection are sensitive to small perturbations of the experimental conditions. One method to “stabilize” variable selection explored in this issue is to use several “bootstraps” (Bi et al. 2003).

monotonic

dichotomous

# 6. Conclusion

Our experiments show that transforming features using RBMs can improve both results of similarity learning with gradient ascent and Support Vector Machines on music audio. ####################

The study makes several contributions. First, the application of machine learning approaches to feature processing where very sparse ####. Second, show potential applicability of statistical techniques in rich financial data where interpolation is not a feasible solution. #### #### Third, the models proposed from grid search prove to be conducive to better fitting of optimization parameters, which especially when combined with consistent models features are most useful. #### #### feature selection are related to model specification and algorithm selection, comprehensibility of complex models become important .. Especially, as feature embedding to be fed into a more complex deep neural network architecture.

Future research is encouraged for different shapes of data. While traditional financial data sets are mostly structured, recent advancements in algorithms have shown that unstructured data can effectively identify previously unknown patterns and relationships. A few examples financial institutions can employ include speech-to-text (STT) data gleaned from customer calls, customer location information (GIS) tracked from mobile apps as well as off-line purchases, and 2D image data generated from customers’ social media accounts. The increasing heterogeneity of input data will pose a challenge in computation and model interpretation. Further studies on constructing and evaluating features would help efficient discovery of knowledge without bold assumptions on variable distributions.

# REFERENCES

Ban, G.-Y., and Rudin, C. 2018. “The Big Data Newsvendor: Practical Insights from Machine Learning,” *Operations Research* (67:1), pp.

Ranzato, M., Boureau, Y., and LeCun, Y. 2007. “Sparse Feature Learning for Deep Belief Networks,” in *Proceedings of Neural Information Processing Systems*, pp. 1-8.

Hinton, G. E., Osindero, S., and Teh, Y.-W. 2006. “A Fast Learning Algorithm for Deep Belief Net,” *Neural Computation* (18), pp. 1527–1554.

Kohavi, R., and John, G. H. 1997. “Wrappers for Feature Subset Selection,” *Artificial Intelligence* (97), pp. 273-324.

Tibshirani, R. 1996. “Regression Shrinkage and Selection via the Lasso,” *Journal of the Royal Statistical Society: Series* *B* (58:1), pp. 267-288.

Zou, H., and Hastie, T. (2005) “Regularization and Variable Selection via the Elastic Net,” *Journal of the Royal Statistical Society: Series* *B* (67:2), pp. 301-320.

Huang, X., Song, Q., Yang, F., and Hu, X. 2019. “Large-Scale Heterogeneous Feature Embedding, in Proceedings of Association for the Advancement of Artificial Intelligence, pp.

López de Prado, M. (2018) *Advances in Financial Machine Learning*, Wiley.