Overdue Invoice Forecasting and Data Mining

by

Weikun Hu

Bachelor of Science in Mathematics and Statistics
University of Washington, 2014

Submitted to the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degree of

Master of Science in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2016

Author: .... Signature redacted ....
Department of Civil and Environmental Engineering
May 18, 2016

Certified by: .... Signature redacted ....
David Simchi-Levi
Professor of Civil and Environmental Engineering
Thesis Supervisor

Certified by: .... Signature redacted ....
Heidi Nepf
Donald and Martha Harleman Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

Overdue Invoice Forecasting and Data Mining

by

Weikun Hu

Submitted to the Department of Civil and Environmental Engineering
on May 18, 2016, in partial fulfillment of the requirements for the degree of
Master of Science in Transportation

Abstract

The account receivable is one of the main challenges in the business operation. With poor management of invoice to cash collection process, the over due invoice may pile up, and the increasing amount of unpaid invoice may lead to cash flow problems. In this thesis, I addressed the proactive approach to improving account receivable management using predictive modeling.

To complete the task, I built supervised learning models to identity the delayed invoices in advance and made recommendations on improving performance of order to cash collection process. The main procedures of the research work are data cleaning and processing, statistical analysis, building machine learning models and evaluating model performance. The analytical and modeling of the study are based on the real-world invoice data from a Fortune 500 company.

The thesis also discussed approaches of dealing with imbalanced data, which includes sampling techniques, performance measurements and ensemble algorithms. The invoice data used in this thesis is imbalanced, because on-time invoice and delayed invoice classes are not approximately equally represented. The cost sensitivity learning techniques demonstrates favorable improvement on classification results.

The results of the thesis reveal that the supervised machine learning models can predict the potential late payment of invoice with high accuracy.

Thesis Supervisor: David Simchi-Levi

Title: Professor of Civil and Environmental Engineering

Acknowledgements

First and foremost, I would like to thank my advisor Prof. David Simchi-Levi. He has offered invaluable support, guidance, and knowledge over the last two years for which I am truly thankful.

I would also like to show gratitude to my friend and co-worker, Peiguang Hu, who guided me through the research project during the first year.

Most importantly, I would like to thank all my friends who helped me get through two years of graduate school.

Finally, to my parents, thank you for your love and support throughout this journey.

Contents

List of Figures

List of Tables

1. Introduction

1.1 Context and Significance

The unpaid invoice is one of the main challenges in the operation of the firm. With poor management of invoice to cash collection process, the unpaid invoice may pile up and cause issues in the business. In other words, the increasing amount of unpaid invoice may lead to cash flow problems in the firm. In the business management, collecting unpaid invoice is a very tedious works, and it is reluctant for most of the firms to do the collection.

Firms usually apply cash methods of accounting to deal with the overdue invoice issues. By using these methods, the firm can deduct the amount of accounts receivable, and maintain the sustainable business operation.

In this research work, we proposed the predictive modeling as a proactively preventive management method in the account receivable management. By building supervised model using machine learning algorithms, we could detect the over due invoices in advance, and help the firms to improve their performance in invoice to cash collection process.

1.2 Objective and Research Question

In this project, we studied the payment behavior of invoices for the customers of a technology firm. Instead of working on the payment pursing for overdue invoices in the traditional way, we are interested in detecting the potential delayed invoice in advance using innovative analytics methods. The proactive approach of forecasting with classification model is able to provide recommendations on firm's management on invoice to cash collection process.

The account receivable is one of the main challenges in invoice processing. The objective of the project is the build supervised learning model to identity the delayed invoices in advanced and improve the order to cash process.

Building supervised learning models to forecast overdue invoice in advance can help firm to have a better understanding of the unpaid invoice and related customers who always have late payments on invoices, and be prepared for the late invoices.

The approaches of the project are in the following steps:

1. Data cleaning and pre-processing
2. Statistical analysis and feature selection
3. Building supervised learning models with training data
4. Test and evaluate the performance of classification models

2 Literature Review

2.1 Data Mining and Business Analytics

Data mining and business analytics techniques have been widely used in the finance industry in recent years. The advanced statistical techniques are able to help firm to perform analytical work on large data sets with high efficiency, high productivity, high quality, and relatively low cost.

The machine learning approaches has been successfully implemented in several fields of business. In the literatures of the related topics, many researchers have conducted experiments on fraud detection and credit risk forecast with real-world data from firms and companies.

Fraud detection

Fraud detection has been a popular topic in the research of machine learning over the past ten years. Many researchers have been working on the to implement the learning algorithms into the fraud analysis. The research works by Phua *et al.* have discussed four major approaches of data mining on identifying the suspicious instances of transactions (Phua et al. 2010).

- Supervised learning on labeled data: the most commonly used algorithms are neural network and support vector machine.
- Hybrid supervised learning: the algorithms introduced in the hybrid-supervised learning are weighted Bayesian network (Ormerod et al. 2003), and optimal weighted attributes with k nearest neighbor (He, Graco, and Yao 1998).
- Semi-supervised learning on non-fraud data includes auto-associative neural network (Aleskerov, Freisleben, and Rao 1997), decision trees with Boolean logic function (Kokkinaki 1997), etc.

- Unsupervised learning with unlabeled data, which includes unsupervised neural network (Dorronsoro et al. 1997), outlier detection (Yamanishi et al. 2004), etc.

Learning from previous research work on fraud detection, we found the advanced learning methods have been improving the detection process significantly and reach the cost efficient on the operation level (Phua et al. 2010).

Credit Risk

Classification tree modeling has been used in building customer credit risk model with a high accurate in credit forecasting (Khandani, Kim, and Lo 2010).

Unsupervised learning model has been discussed in credit risk score card construction (Correa et al. 2012). In the paper, researchers use the cluster analysis as part of learning algorithm in building credit risk score card and compare the performance of clustering with other conventional methodology. They find the cluster analysis of distance methodology has much higher predictive accuracy compare to logics regression and MLP neural network algorithms.

2.2 Analytical Research on Invoices

Over the past years, the advanced statistical techniques and analytical analysis have been used to improve invoice process.

The implementation of Lean Six Sigma phase's road map in the invoice to cash collection process has been discussed in a recent study. The research showed the clearly define and measure of the problem is able to make significant improvement on the invoicing process. The approaches of the project was to reduce the throughput and processing time of invoicing, and improve the collection efficiency and reduce the late payment of invoice (Erdmann, Groot, and Does 2010).

Markov model has also been used to evaluate invoice processing. The approach is used to detect and rank bottlenecks to prioritize of process improvement. (Younes et al. 2015)

Cash inflow curve analysis is also an approach to invoicing forecasting. An intelligent system that based on case-based reasoning (CBR) has been used in cash flow forecasting (Simić, Simić, and Svirčević 2011). In their study, they proposed the concepts of invoice curve and cash inflow curve analysis. By knowing the saturation point of cash inflow in the future, the prediction of invoiced activities can be sufficient reliably.

Using machine-learning models to forecast the payment process of invoice has been developed recently, but there are few literatures have been presented. Supervised learning has been implement in the improvement of invoice to cash collection process (Zeng et al. 2008). Their models show high prediction accuracy on delay payment of invoices, which provides useful framework for the machine learning study on invoice collection process.

2.3 Imbalance Data in Predictive Modeling

In data mining field, several challenges have become pronounced in classification problem. The imbalance nature of data sets is one of the problems that have been widely discussed. The imbalance data refers to data sets that composed by a normal group with majority of the instances and an abnormal (interesting) group that only contains a few instances. There are many typical problems have issue with imbalanced data sets, such as fraud detection and disease diagnostic (Chen 2004).

In the learning process, the model is more likely to capture the features from prevalent group and less focus on rare instances. Therefore, most of classifiers would not have a good performance on these types of imbalance dataset, and it is very difficult to detect the under-represent class (Ling and Li 1998). Ignoring the under represented class leads to substantial consequences both in model estimation and accuracy evaluation on estimated model (Menardi and Torelli 2014).

Imbalance class distribution is an important challenge in data mining field, and much of research works have addressed the approaches to improve prediction performance of the classification algorithm. The popular application research topics in the imbalanced dataset are in disease diagnostics and credit card fraud detection. There are two major methods have been widely discussed in imbalanced datasets classification, the sampling techniques and appropriate performance measurements (Maimon and Rokach 2005).

2.3.1 Sampling techniques

Traditional sampling methods

Sampling techniques are used to balance datasets, and traditional methods are over-sampling the minority class, under-sampling the majority class, or a combination of both methods. Various studies have discussed the trade off between the over-sampling and under-sampling. And some of the research found that under-sample of the majority class would lead to a better classification model than the over sample of the minority class (Chawla et al. 2002). The random over-sampling may bring over fitting problem, while under-sample could eliminate certain significant instances in the datasets (Maimon and Rokach 2005).

There are a few refined methods that could be used to enhance the performance of binary classification algorithm on imbalanced dataset.

Synthetic generation methods

Synthetic Minority Over-sampling technique (SMOTE)

Over sampling with replacement technique is able to increase the information on minority class in the training set and keep the two classes in the equal data size. However, this method not generally works well on all learning algorithms. For example it does not make a significant improvement on Naïve Bayes algorithm (Ling and Li 1998).

Due to the limitation of sampling methods, the advanced methods like Synthetic Minority Over-sampling technique (SMOTE) has been proposed in solving imbalance dataset problem (Chawla et al. 2002).

The SMOTE is a sampling method that over-sampling the minority class by generating synthetic examples. The traditional minority class over-sampling method is sampling with replacement. In SMOTE approach, every instance in the minority class was taken into sampling and synthetic instances along the line segments of the k minority class nearest neighbours was introduced. Neighbours from k nearest neighbours are randomly chosen based on the amount of over-sampling needed. Compared with over-sampling with replacement, SMOTE algorithm effectively identifies more general decision regions in the feature space for the minority group. Given more general regions in the feature space, the features of minority class that used to be bounded by the majority group could be well learned by the classifiers.

The Figure 2-1 presents the comparison between the methods of minority over-sampling with replacement and SMOTE using the mammography dataset (Chawla et al. 2002).

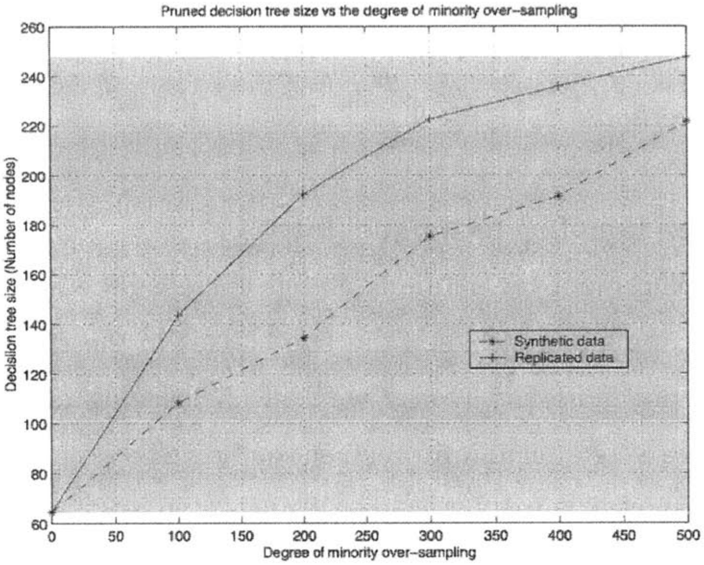


Figure 2-1 ROC Curve

Random Over-Sampling Examples (ROSE)

Random over-sample examples (ROSE) is a smoothed bootstrap-based technique that used to mitigate the effect of extreme imbalanced distribution of classes. The ROSE method benefits the processes of model estimation and assessments. The ROSE method is applicable on both continuous and categorical dataset, and it generates synthetic examples from the conditional density estimate of the two classes.

2.3.2 Performance measurements

Confusion matrix

Confusion matrix is a table layout that allows visualization of the performance of the supervised learning algorithm in the field of machine learning. In the table, the column of the matrix represents the instances in a predicted class, and the rows of the matrix represents the instances in an actual class. The confusion matrix is type of contingency table that has two dimensions, i.e. predicted and actual, and both of the dimensions have identical sets of classes. For example, a classifier has been trained to distinguish between two groups, diseased and normal groups, and the confusion matrix would summarize the classification results of the algorithm for further inspection.

Table 2-1 is an example of confusion matrix for binary classification.

		Predicted Class	
		Normal	Diseased
Actual Class	Normal (Positive)	True Positive	False Negative (Type II error)
	Diseased (Negative)	False Positive (Type I error)	True Negative

Table 2-1 Confusion Matrix

Terminology of confusion matrix:

True positive (TP): actual normal class is correctly classified as normal class.

True negative (TN): actual diseased class is correctly classified as diseased class.

False positive (FP): diseased class is incorrectly labeled as normal class.

False positive (FP): normal class is incorrectly labeled as disease class.

True positive rate (TPR)/ Sensitivity/ Recall:

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

True negative rate (TNR)/ Specificity (SPC):

$$SPC = \frac{TN}{N} = \frac{TN}{FP + TN}$$

Precision/ Positive predictive value (PPV):

$$PPV = \frac{TP}{TP + FP}$$

Accuracy (ACC):

$$ACC = \frac{TP + TN}{P + N}$$

F1 score: harmonic mean of precision and sensitivity:

$$F1 = \frac{2TP}{2TP + FP + FN}$$

The cut-off value selection is based on the purpose of the test. The accuracy is a common approach on evaluating the prediction accuracy of the algorithm for balanced datasets, when the cost of misclassification for both classes is equal. For imbalanced data set, the accuracy is not a proper metric to measure the performance of the algorithm, because only a small percentage of the data lies in the interesting group. And the cost of misclassifying an instance in the underrepresent group is much higher than the misclassification in dominant group. For instance, a classification model could classify all instances as normal with a high accuracy, but it fails in detecting the rare diseased class (Drosou et al. 2014). The ROC curve diagnostics would be an appropriate metric to measure the classification results for imbalanced datasets. The sensitivity (True Positive Rate) and specificity (True Negative Rate) can be used to effectively on evaluating the performance of the classifier for the skewed data. For example, if the classifier is used to confirm a disease, then the higher specificity is preferred (Chawla et al. 2002).

ROC curve analysis

When consider the classification results of a specific statistical model that is used to classify the dataset into two groups, for example, one group is diseased and the other group is normal, the results of the algorithm is hardly a perfect separation between the two groups. In realistic, the distributions of the two groups always have an overlap, and one group is closer to 0 and the other group is closer to 1. The overlap represents the area that the test cannot distinguish one group from other group. The Figure 2-2 is an example of the results.

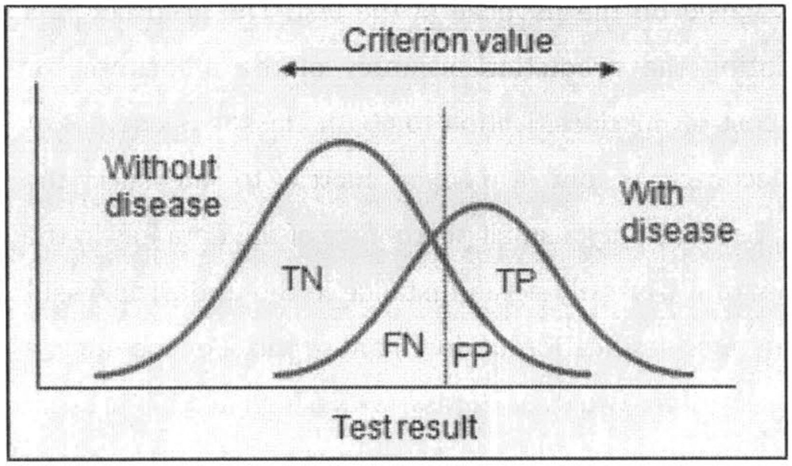


Figure 2-2 Classification Results

For every possible value of cut-off points or criterion value, ranging from 0 to 1, certain fraction of data is misclassified into opposite group. For example, we select the cut-off point, indicated by the dash line, some instances in the normal group on the left are classified as diseased, and some instances in the diseased group are classified as normal. The position of the point determines the true positive (TP), true negative (TN), false positive (FP), and false negative, and the objective of cut-off point selection is to minimize of the erroneous types of the results.

In default classification models, the threshold is usually set to be 0.5. However, this value of cut-off point may not work well in imbalance datasets. By using 0.5 as a threshold in skewed datasets, few instances in testing sets is labeled as minority groups. Using a smaller cut off points could improve the misclassification problem on minority group, because it is equivalent to increase the misclassification cost for the class. The ROC curve could be used for accuracy measurement of the model (Menardi and Torelli 2014).

The performance of a classifier, or the accuracy of a classifier to distinguish two classes of data is measured by the Receiver Operating Characteristics (ROC). ROC curve analysis is an intuitive way to examine the performance of a classifier, and it also used to compare the performance among two or more classification models. It is a curve that plots the true positive rate (Sensitivity) against the false positive rate (1- Specificity) for different possible values of cut-off points. The curves visualize the trade-off between sensitivity and specificity using graphs. For example, an increase in the sensitivity would lead to a decrease in specificity. The points on the curve represent pairs of true positive rate (TPR) and true positive rate (FPR) under specific thresholds. The Figure 2-1 is an example of ROC curve.

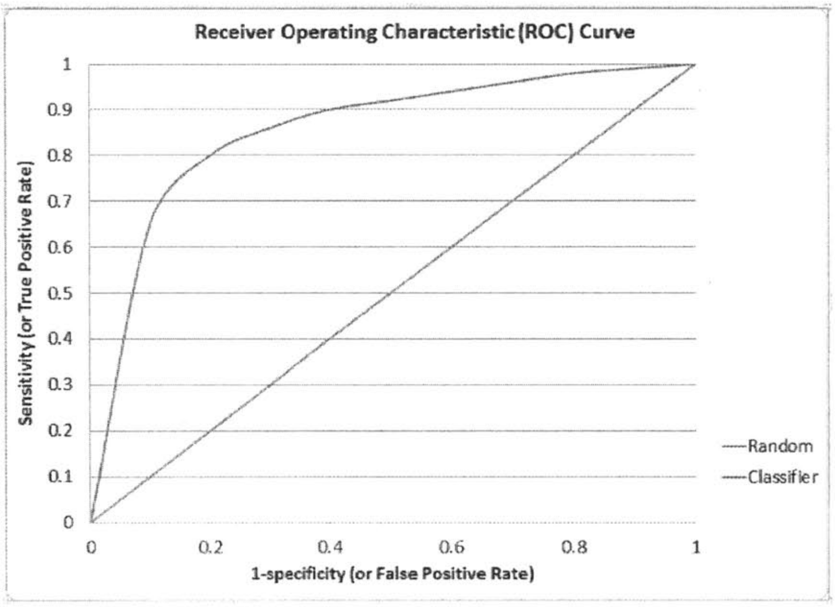


Figure 2-3 ROC Example

The area under the curve (AUC) quantifies the ability of the classifier to distinguish between two groups (diseased/normal). AUC represents an area within the unit square, and the value of AUC is between 0 and 1. The random guess of the classification would generate the diagonal line connect (0,0) and (1,1), and the perfect classifier should go through the upper left corner point (0,1), which gives

100% on TPR (Fawcett 2006). A realistic ROC should lies between the diagonal line and upper left corner point, and the distance between ROC and diagonal line is an indicator of the performance of classifier. In other words, the closer the ROC to the upper left corner point, the higher the accuracy of the classifier (Zweig and Campbell 1993).

3 Data Overview and Data Process

The analysis works in the following sections are based on the invoice data sets from a fortune 500 company, providing technology service. The invoice data is in monthly frequency, and we were given six months of data, ranging from November 2014 to April 2015. Figure 3-1 shows the overview of all invoices data.

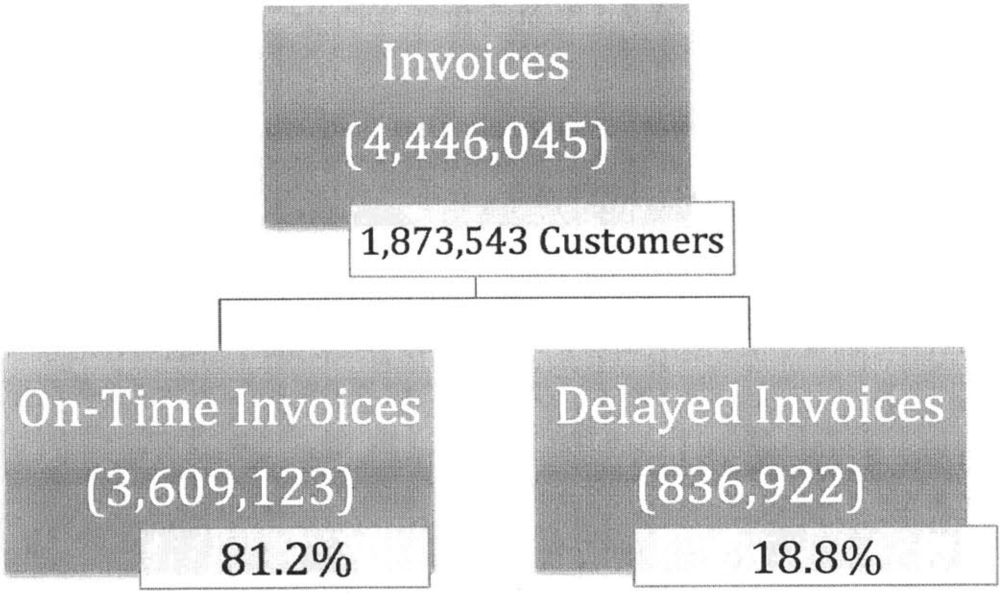
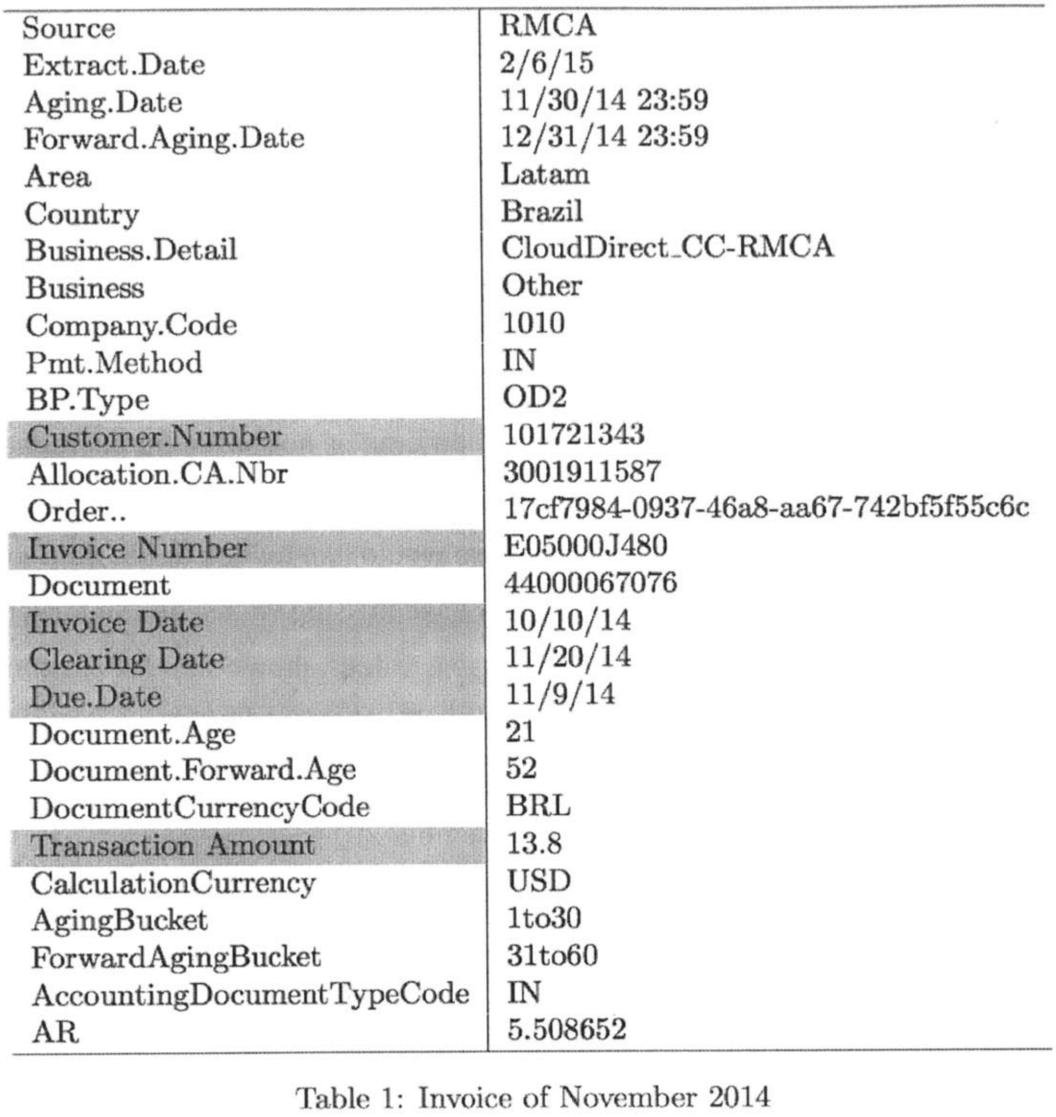


Figure 3-1 Data Summary

3.1 Raw data and data formats

The invoice raw data sets contain monthly invoices from the firm divided into four regions: APAC (Asia and Pacific), EMEA (Europe, the Middle East and Africa), NA (North America), and LATAM (Latin America). On average, the firm has 700,000 invoices every month. Each row of the data sets is an invoice, and each invoice has 59 features that provides invoice information such as customer number, order details, transaction amount, invoice date, etc. After initial inspection, we removed variables that do not provide meaning information, such as *"FWD_*"*. Then we eliminate the variables with a lot of missing value and variables that have single value, such as *"purchase order number"*. After preliminary selection, we obtained the simplified features of invoice in the Table 3-1. For now, we have 28 features for invoices. Then we filter out feature have missing value or unique value.



Source	RMCA
Extract.Date	2/6/15
Aging.Date	11/30/14 23:59
Forward.Aging.Date	12/31/14 23:59
Area	Latam
Country	Brazil
Business.Detail	CloudDirect_CC-RMCA
Business	Other
Company.Code	1010
Pmt.Method	IN
BP.Type	OD2
Customer.Number	101721343
Allocation.CA.Nbr	3001911587
Order..	17cf7984-0937-46a8-aa67-742bf5f55c6c
Invoice Number	E05000J480
Document	44000067076
Invoice Date	10/10/14
Clearing Date	11/20/14
Due.Date	11/9/14
Document.Age	21
Document.Forward.Age	52
DocumentCurrencyCode	BRL
Transaction Amount	13.8
CalculationCurrency	USD
AgingBucket	1to30
ForwardAgingBucket	31to60
AccountingDocumentTypeCode	IN
AR	5.508652

Table 1: Invoice of November 2014

Table 3-1 Sample of Invoice

The features that we used in model and related calculation are highlighted in Table 3-1, which includes Customer Number, Invoice Number, Invoice Date, Clearing Date, Due Date, Transaction Amount.

3.2 Data Process

3.2.1 Data Cleansing

To prepare for the analytical part of the work, we would like to combine the datasets in to appropriate groups, and clean the data sets and transform them into the desired formats. First, we selected the closed invoices from the whole data sets, because these invoices provide us with the delay information that we need in the statistical modeling part. If invoices are closed, then the invoices are paid. And we would know whether the invoices have been paid on time or not based on payment date information. Secondly, for each month, we combine the invoices from four regions (APAC, EMEA, NA and LATAM), because we would like to look at the data in monthly basis. Then, there are many invoices that have missing values on the target variables, such as "customer number". The missing values shown as "NA" in the data sets, and we remove the instances with missing value. The summary of invoices data shows in the Table 3-2. For example, the invoices data sets contain 409, 158 invoices from 268, 622 customers in November 2014.

Time	Number of Invoices	Number of Customer
November, 2014	409, 158	268, 622
December, 2014	811, 818	422, 212
January, 2015	687, 046	396, 924
February, 2015	754, 912	433, 140
March, 2015	904, 510	567, 342
April, 2015	878, 601	590, 059

Table 3-2 Invoices Summary Table

3.2.2 Data Processing

3.2.2.1 Additional Features

As mentioned in section above, there are features that we are interested in the invoices information. And we would like to create additional information based on existed features and add them into the model.

- End of month indicator: whether the invoice due at month end or not. It is a binary indicator. If the invoice is due in last three days of the month, the value of end of month indicator is 1, and 0 other wise. For example, the invoice in Table 3-1 due on November 9th, 2014, and the value of indicator is 0 for this invoice.

- Second half of month indicator: whether the invoice due at second half of the month or not. If the due dates of invoice is after the 15th of the month, the value of end of month indicator is 1, and 0 other wise. For example, the invoice in Table 3-1 due on November 9th, 2014, and the value of second half of month indicator is 0 for this invoice.

- Delay Days: the delay days are the differences between *clearing dates* and *due dates* of the invoices, which is subtracting due dates from clearing dates. If "delay days" is positive, the invoice is not paid on time, and it is defined as delayed. If "delay days" is non-positive, the invoice is paid on time, and we say it is non-delay invoice. For example, the payment of invoice in Table 3-1 was due on November 9th, 2014, and was cleared on November 20th, 2014. The delay days of this invoice is 11 days, and the invoice is classified as delayed invoice. For invoices have non-positive value of delay days, we set the delay days to be zero.

- Payment term (buffer): the difference between invoices' *due date* and *invoice date,* and it is calculated by subtracting *invoice date* from *due date*. For example, the invoice in Table 3-1 is issued on October 10th, and the payment of the invoice is due on November 9th. S0, the payment term of this invoice is 30 days.

4 Preliminary Analysis

4.1 Problem Formulation

The objective of our research work is to solve the following challenge:
Given a new instance with a set of features, predict whether the payment of invoice is delayed or not.

The general approach for the delayed invoice detection problem is building classification models. There are two outcomes for each invoice, and the new invoice will be labeled as either "on time" or "delayed". The supervised learning models built based on features of invoices would be able to classify the data into an appropriate outcome group.

4.2 Dependent and Independent Variables

The independent variables of the classification models are selected from features attached to the invoice, which have high performance on distinguishing instances from different outcome groups. These features are representative and facilitate generalization.

The independent variables have two levels.

- Invoice level
- Customer level

Adding customer level features to invoice would provide useful history information of customer behaviors, which could improve the accuracy of predictive models. The logic of combining features at both invoice and customer level shows in Figure 4-1.

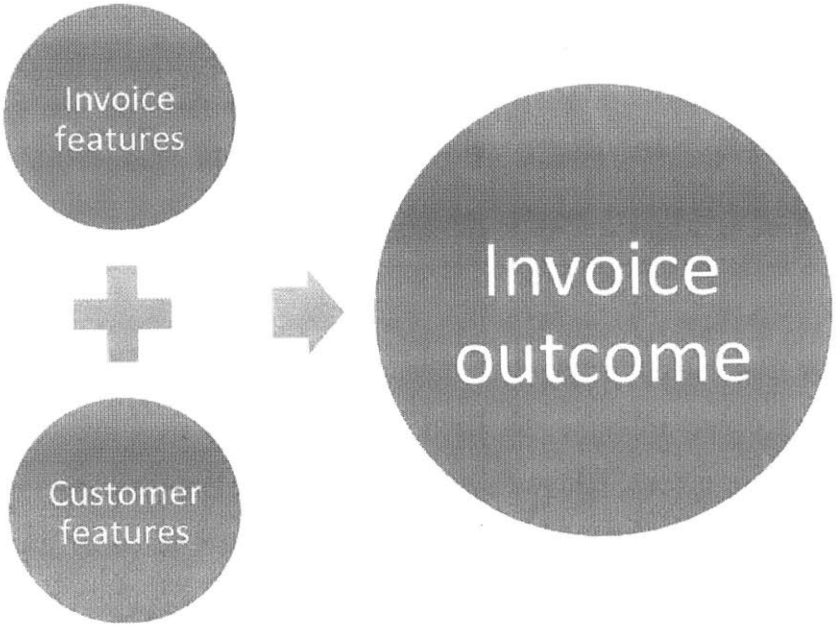


Figure 4-1 Invoice and Customer Features

For invoices level, we have features directly related to individual invoices. For example, the transaction amount of invoice is one of the independent variables from invoice level. The features at invoices level are listed in Table 4-1.

	Name of Variable	Description
1	Transaction Amount	--
2	End of month indicator	Whether the invoice due at month end or not
3	Second half of month indicator	Whether the invoice due at second half of month or not

Table 4-1 Independent Variables at Invoice Level

For features on customer level, we aggregate invoices that belong to same customer and create a customer profile that could be attached to individual invoices based on the *Customer Number*. For example, we calculate the total number of invoices for each customer. The features at customer level listed in Table 4-2.

	Name of Variable	Description
1	Number of Invoice	Total number of invoices
2	Number of Delayed Invoice	--
3	Total Amount of Invoice	Total transaction amount of invoice
4	Average Amount of Invoice	--
5	Total amount of delayed invoices	--
6	Average amount of delayed invoices	--
7	Delay ratio	Ratio of 1 and 2
8	Amount ratio	Ratio of 3 and 4
9	Average payment term	The difference between invoices' *due date* and *invoice date*
10	Average delayed days	--

Table 4-2 Independent Variables at Customer Level

The binary outcome of invoice has defined by the delay days as the following:

- On time: the invoice is paid before or on due date
- Delayed: the invoice is paid after the due date

The dependent variable of our classification model is the payment status of invoices, i.e. the outcome of the invoice. Based on the due date and clearing date, we are able to determine whether the invoice is paid on time or not.

4.3 Metrics used to evaluate models

For the classification problem of invoices, appropriate metrics should be selected to evaluate the performance of the predictive models. In our research project, we use confusion matrix and related calculation to measure the prediction results of supervised learning algorithm. Confusion is a table layout that presents the classification results of the predictive models, and Table 4-3 shows the confusion matrix used in our analysis on delayed invoices.

		Predicted Classes (Predicted by the classification model)	
		On Time	Delayed
Actual Classes	On Time	**True Positive (TP)**	**False Negative (FN)**
	Delayed	**False Positive (FP)**	**True Negative (TN)**

Table 4-3 Confusion Matrix for Invoices

In Table 4-3, the column of the matrix represents the instances in a predicted class, and the rows of the matrix represents the instances in an actual class. The significant terms and notations in the confusion matrix are defined in the following.

- True positive (TP): on time invoice is correctly classified as on time class.
- True negative (TN): delayed invoice is correctly classified as delayed class.
- False positive (FP): delayed invoice is incorrectly labeled as on time class.
- False positive (FP): on time is incorrectly labeled as delayed class.

There are a few statistical measurements of prediction accuracy that derived from the confusion matrix, such as accuracy, precision, prevalence, etc.

We would combine the accuracy and specificity measurements to evaluate the performance of the model of invoice classification. The accuracy is the proportion of the true results (TP and TN) among the total number of instances examined by predictive model. Equation 4-1 shows the formula of accuracy calculation. The specificity (also called true negative rate) measures the proportion of delayed invoices that are correctly identified as delayed invoices. Equation 4-2 shows the mathematical formula of specificity. The main objective of the project is to identify delayed invoices, so we are more interested in delayed class compare to the on time class. Therefore, we use accuracy and specificity to measure the performance of machine learning algorithm, and we are more interested in the accuracy measured by the specificity. Misclassification of delayed invoices is more costly compare to on time invoices.

Equation 4-1 Accuracy

$$ACC = \frac{TP + TN}{P + N}$$

Equation 4-2 Specificity

$$SPC = \frac{TN}{N} = \frac{TN}{FP + TN}$$

4.4 Statistical Analysis

In this section, we would use the invoices in November 2014 as an example of our statistical analysis on the delay of invoice.

Figure 4-2 is a pie chart shows the percentage of delayed invoices among total invoices in the month of November. From the graph, we find the delayed invoice is only 28% of total invoices. The delayed class is the minority class in the invoice data sets, and the data set is imbalance in terms of the class distribution.

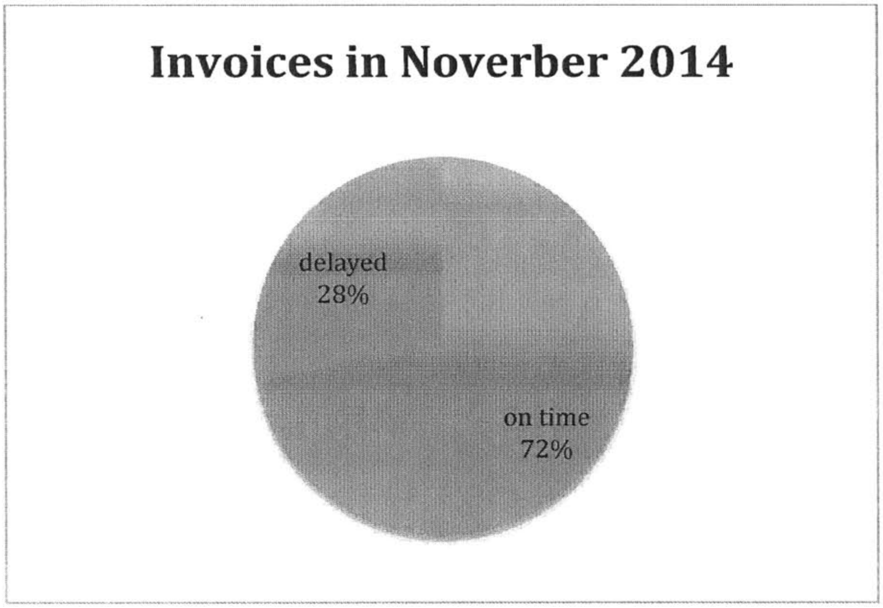


Figure 4-2 Pie chart of Invoices in November 2014

Then we would like to analysis the distribution of delay days of invoices using the Figure 4-3, which is the histogram of the delay days of each invoices. Seeing from the graph, we find the days of delay are concentrated within the 30 days period, and there are some extreme cases that the delay is more a year.

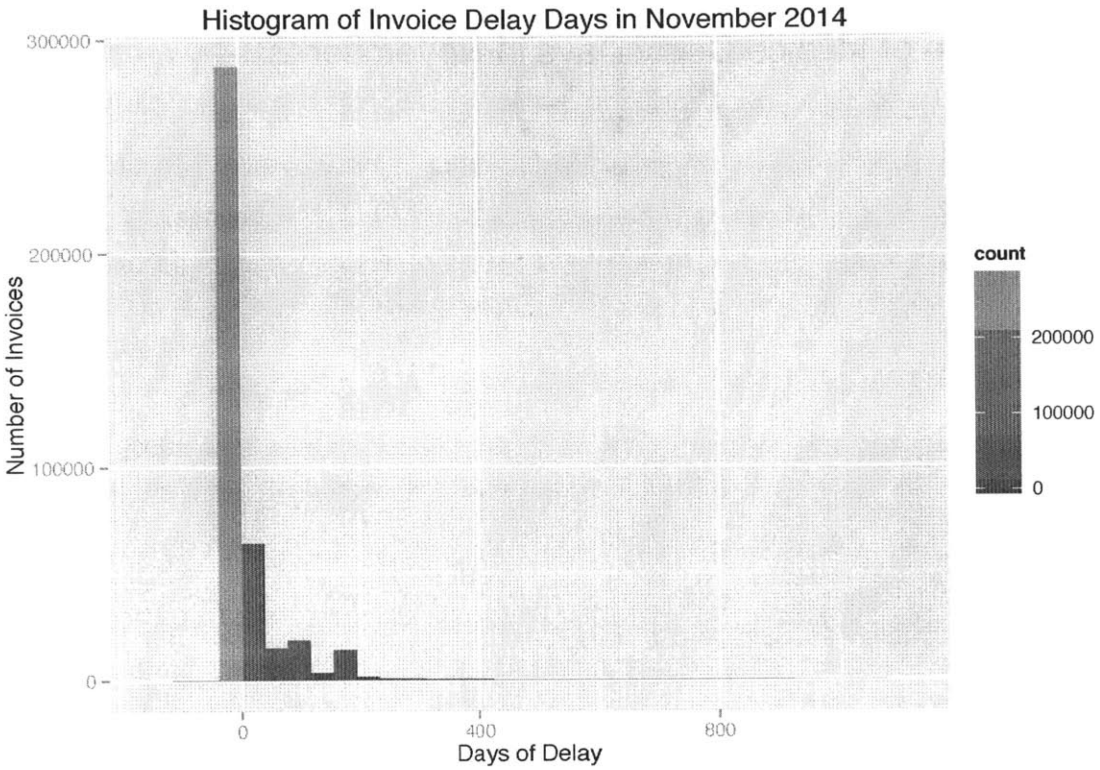


Figure 4-3 Histogram of Delay Days

Then we want to study the pattern of the delayed invoice. In the previous part, we defined the delayed invoice, which is the invoice that has positive delay days. In Figure 4-4, we show the distribution of delay days for delayed in voices only, and remove some extreme invoices that have larger delay days. From the graph, we find most delayed invoices have reasonable length of delay, which is the delays are within one month. However, there are certain amount of invoices, have delayed for more than half of the month. We need to pay attentions to these invoices and try to find if there any common features share by these invoices.

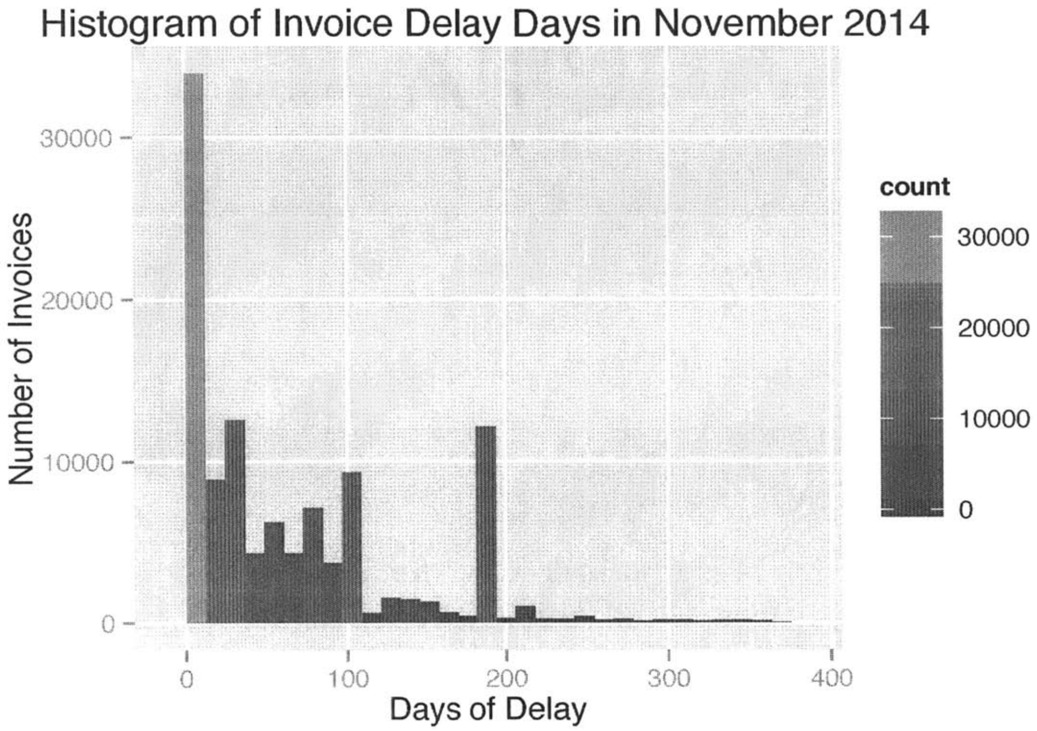


Figure 4-4 Histogram of Delay Days for Delayed Invoice

Based on the analytical results on data sets from November 2014, we found that the delayed invoice class is a minority class compared to on time invoices class. And the length of delays is centered within 30 days period, and with some longer delays that ranging around half year length.

Then we conduct the same analysis on other month data and summarized the delayed invoices information for six months invoice data in Table 4-4.

Time	# of invoices	# of delayed invoice	% of delayed invoices
November, 2014	409158	114109	27.9%
December, 2014	811818	269947	33.3%
January, 2015	687046	147875	21.5%
February, 2015	754912	147406	19.5%
March, 2015	904510	96590	10.7%
April, 2015	878601	60995	6.9%

Table 4-4 Delayed invoice summary

4.5 Skewness of Class Distribution

The primary objective of the project is to detect the potential delayed invoices. The invoices are grouped into two classes based on their payment pattern: invoice paid on time and invoice not paid on time, i.e. non-delayed invoices, and delayed invoices. We would discuss the classification and problem formulation of the delay invoices prediction in details in the following sections.

Seeing from Table 4-4, we found the distribution of two classes, i.e. non-delayed and delayed, are skewed. In general, majority of the observations lies in the non-delayed group, and less than 40% of data are delayed. The data sets are extremely skewed for some months. For example, the minority class in April 2015 only has 7% of instances. In that case, the data sets are imbalanced, and we need to implement techniques mentioned in the preview sections to deal with the imbalanced data sets issues.

5 Classification Models

5.1 Introduction to Machine Learning

In this section, we would give an overview of machine learning procedures and explained the standard processes of predictive modeling. Figure 5-1 is the flow chart of supervised learning modeling. The graph presents the main ideas and typical steps of performing supervised learning.

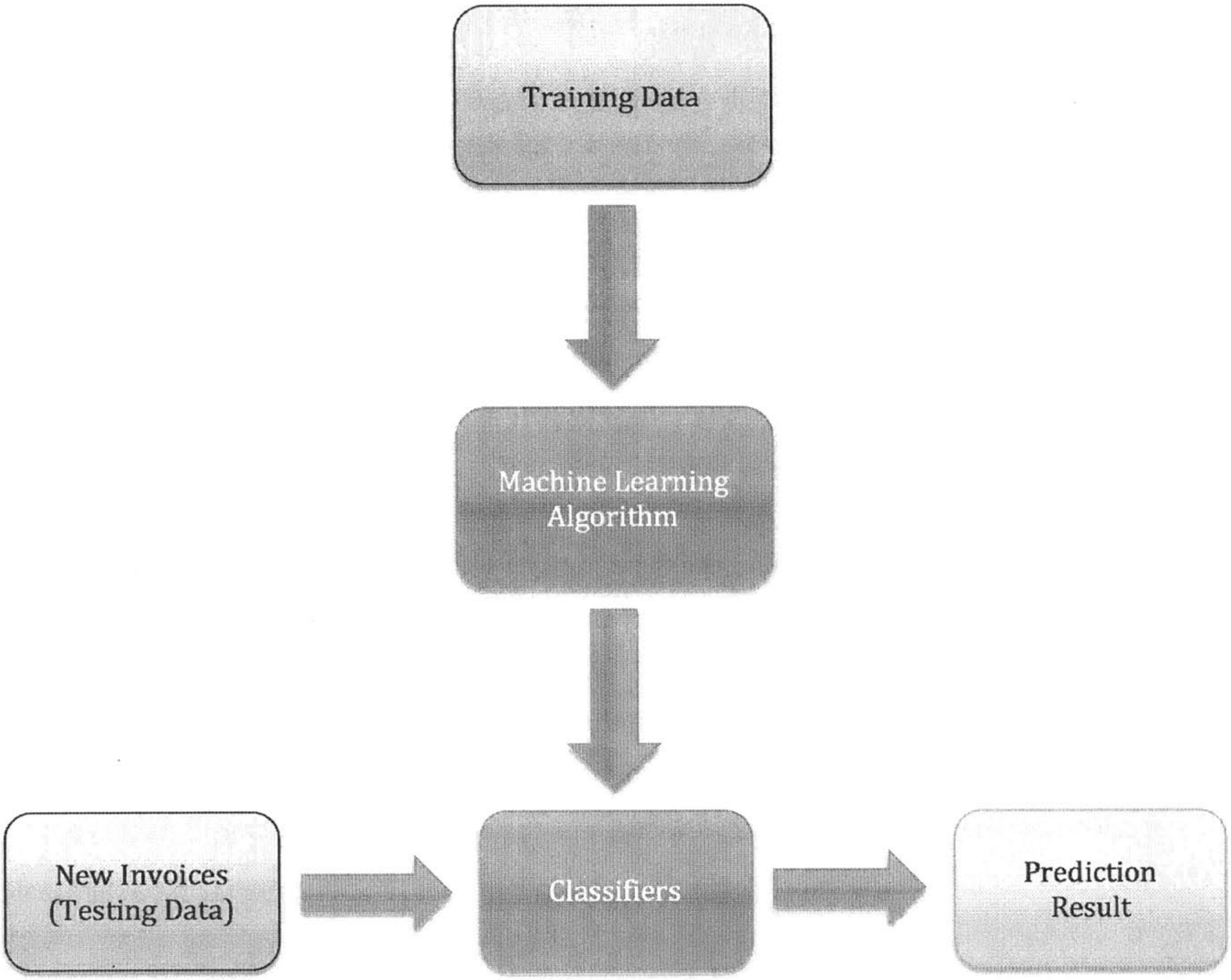


Figure 5-1 Process of Machine Learning

There are two types of learning in the machine learning, the unsupervised learning and supervised learning. The unerpservised learning is using algorithme to group data with similart features into clusters. While the supervised learning is to label instances based on the algorithem. In our problem, we want to detect the delayed invoices, and label the invoices with high chance of delayed given all the historical dealyed invoices features and informtiaon.

5.1.1 Data processing

Before implement the machine learing models, we first sampling data sets and randomly splited the data into training and testing sets in the ratio of 8:2, i.e. 80 percentage of invoice data is used as training sets, and 20 percentage of data is used as testing sets. The training set is used to train the classification model with certain features of invoices, such as transcation amount. The testing set used to measure the performane of classifers.

The classification model is one types of the predictive model, which is able to make prediction on new data based on the historical data. The machine learning algorithme could caputres properties of data in training set and then make prediction on new instances.

5.1.2 Classificaiton models

In the field of supervised machine learning, there are many algorithms that available for use. In the invoice prediction, we selected the following algorithm and implement them on our invoices data.

5.2 Supervised Learning Models

Linear classifier help use to learn the weight or coefficient for each features from training set. Given the input, the output is weighted by the sum of the input.

5.2.1 Logistic regression

Logistic Regression is a fundamental learning algorithm in supervised learning, and it is seen as an analogy for linear regression in classification problem.

Least Square Regression is the most commonly used linear regression model, which is the standard approach to data fitting. The best fit of least square regression model was given by minimized the sum of squared residuals. The least square regression is widely used for modeling, and it could be used as a starting point of the predictive modeling.

The logistic regression is designed to predict the probability of an event. The dependent variable of the model is in finite set of value, and it is usually a binary variable with value of 0 or 1.

An example of the application of logistic regression: given binary dependent variable Y, model the conditional probability as a function of dependent variables x's, and unknown parameters are estimated by the maximum likelihood.

The mathematics behind the logistic regression is the logit function (Equation 5-1). We model the logit ($\log \frac{p}{1-p}$) as a linear function of x. The p in the logit is the probability with value between 0 and 1 (Equation 5-2). The logit transformation make it bounded and with meaningful results.

Equation 5-1 Logit function

$$\log \frac{p(x)}{1 - p(x)} = \beta_0 + x \cdot \beta$$

Equation 5-2 Probability

$$p\ (x) = \frac{e^{\beta_0 + x \cdot \beta}}{1 + e^{\beta_0 + x \cdot \beta}} = \frac{1}{1 + e^{-(\beta_0 + x \cdot \beta)}}$$

The logistic regression could find weights for each feature in the model by showing the coefficients in the model. The signs and magnitude of the independent variables are important in the model interpretation. The positive weights implies the variable is positively correlated with the out come, while the negative weights implies the variable is negative correlated with outcome. The magnitude of the weight indicates the strength of the correlation.

5.2.2 K nearest neighbor

K nearest neighbor (KNN) is a simple classification algorithm and it works well in practical cases. It has been widely used in statistical estimation and pattern recognition field.

The concept of KNN is to predict the class of a new data point using the characteristics of k nearest neighbors in the training set. In other words, the learning algorithm remember the training data, and when making prediction on new data, the model find the nearest data point in training set and return the label associate with the training data point. When picking the k nearest points, the algorithm considers points from training set that are similar to the given point.

The similarity of two data points in KNN is measured by the distance between the points. There are many options for the metric of measurement. Minkowski distance is one of the popular measurements of distance. The formula of Minkowski distance shows in Equation 5-3.

When p is equal to 1, the metric called Manhattan distance (Equation 5-4).
When p is equal to 2, the metric called Euclidean distance (Equation 5-5)

Equation 5-3 Minkowski Distance

$$Dist(x_1, x_2, p) = \left(\sum^{K} abs\ (x_{1k} - x_{2k})^{\ p}\right)^{\frac{1}{p}}$$

Equation 5-4 Manhattan Distance

$$Dist(x_1, x_2, 1) = \sum^{K} abs\ (x_{1k} - x_{2k})$$

Equation 5-5 Euclidean Distance

$$Dist(x_1, x_2, 2) = \left(\sum^{K} abs\ (x_{1k} - x_{2k})^{\ 2}\right)^{\frac{1}{2}}$$

There are other measurements of distance

- Chebyshev distance: measures distance assuming only the most significant dimension is relevant
- Hamming distance: identifies the difference bit by bit of two strings
- Mahalanobis distance: measures distance following only axis-aligned directions

In classification learning, the new data point is classified by the majority vote of its k nearest neighbors. Figure 5-2 gives an example of KNN classification when k equals to 5. The instance x_q is classified as negative, because three out of five of its nearest neighbors are in negative class.

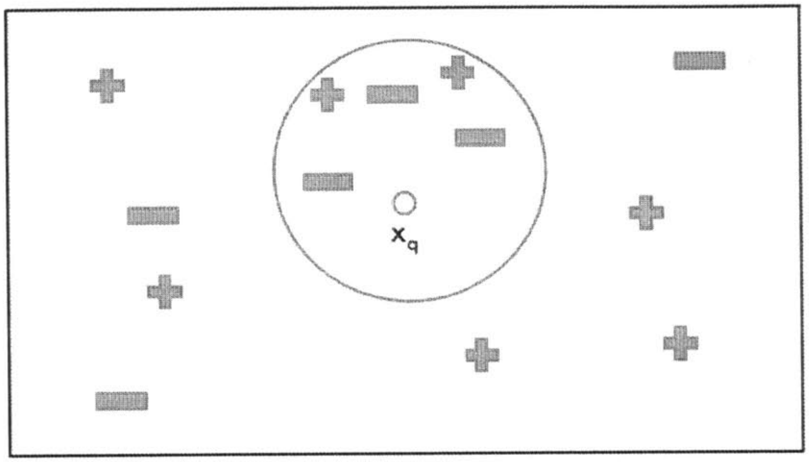


Figure 5-2 KNN for k = 5

The important parameter of the KNN is the k, and we need to specify an integer value of k when applying the algorithm. In general, larger value of k gives a better prediction results, and the optimal k for most of datasets are from 3 to 10 from experiences. Some empirical results showed that square root of number of features could also be an option for value of k. The optimal value of k could be selected by cross-validation methods for specific dataset (Sutton 2012).

There are advantages and disadvantages of the KNN:

For advantage:

- The learning algorithm is simple and fast
- KNN is a non-parametric classifier, and it is helpful when there is no intuition about underlying model

For disadvantage:

- The prediction process might be slow
- Need to store large amounts of training data

5.2.3 Decision trees

Decision tree applies recursive-partitioning method to build prediction models. The models are obtained by recursively partitioning the data space and fitting a prediction model within each partition.

The procedure of classification tree is first asking question at internal nodes, and then answer at leaves. The optimization practice in decision tree is to find the best split. The algorithm designed to find the best split on separating the population. The measurement of separation is calculated by the amount of disorder in leaves, which is represented by entropy. The greedy approach of the algorithm is to find the lowest average entropy for all possible splits. (Introduction to Computation and Programming Using Python)

For large data set, the classification can be huge and complex, and it is important to prune the tree and stop the algorithm at appropriate stage.

The classification tree is in Figure 5-3. The most importance features of invoice in classification model are the transaction amount, average delay days of delayed invoice, and the middle of month indicator. The detailed classifier shows in Table 5-1.

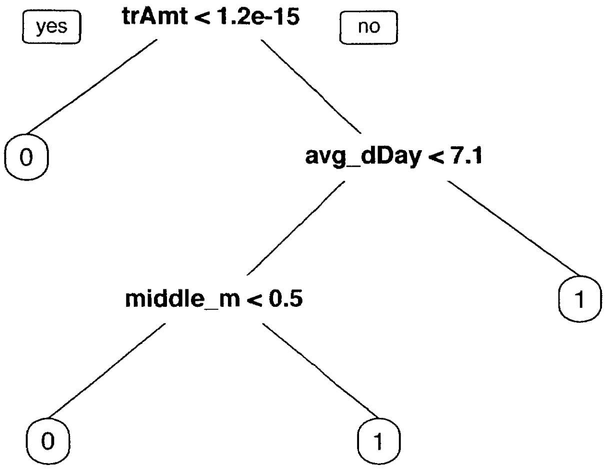


Figure 5-3 Classification Tree

1) root 127267 53769 0 (0.577510274 0.422489726)
2) trAmt< 1.229919e-15 21070 151 0 (0.992833412 0.007166588) *
3) trAmt>=1.229919e-15 106197 52579 1 (0.495108148 0.504891852)
6) avg_dDay_delay< 7.083333 46333 16816 0 (0.637062137 0.362937863)
12) middle_month< 0.5 27235 6693 0 (0.754250046 0.245749954) *
13) middle_month>=0.5 19098 8975 1 (0.469944497 0.530055503) *
7) avg_dDay_delay>=7.083333 59864 23062 1 (0.385239877 0.614760123) *

Table 5-1 Classification Tree Results

5.2.4 Neural Network

Neural Network (NNet) method is inspired by the biological neural network in human brain, which is usually used for abundance of data with little underlying theory. In abalone case, there is no specific theory defines the relationship between age and physical measurements, so it is reasonable to fit the data with Neural Network model.

I first fitted dataset with neuralnet package in R, which could provide multiple hidden layers to build the neural network. However, the neuralnet function was extremely time consuming on abalone dataset. Then I decided to use nnet package to build the model, which only fits single layer neural network. It turned out that a single layer model still provided us with an acceptable performance.

Therefore, we set the hidden layers equal to one based on our model, and then I tuned the size and decay parameters by bootstrapping with 25 reps. The size stands for the number of units in the hidden layer, and the decay stands for the parameter for weight decay. The tuning parameters selected are size equal to 4 and decay equal to 0.064.

The neural network of invoice dataset is showed in Figure 5-4. The color and width of the links between layers are proportion to the direction and magnitude of the weight. The dark color stands for positive weight, and which is excitatory connection in neural biology, and light color stands for negative weight, which is inhibitory connection.

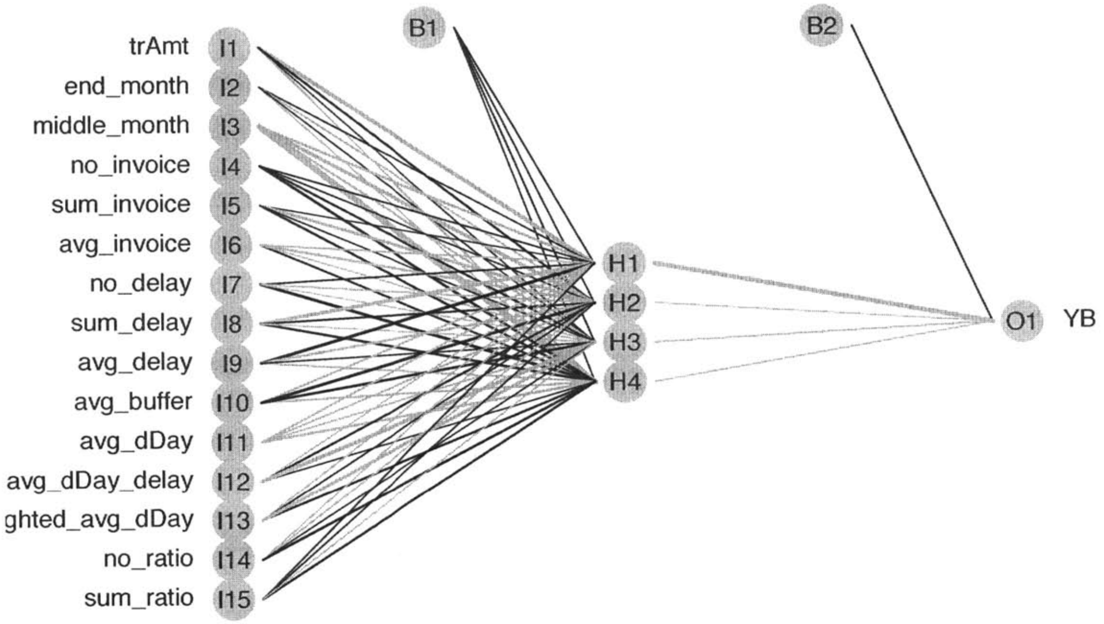


Figure 5-4 Neural Network Plot

6 Ensemble Models

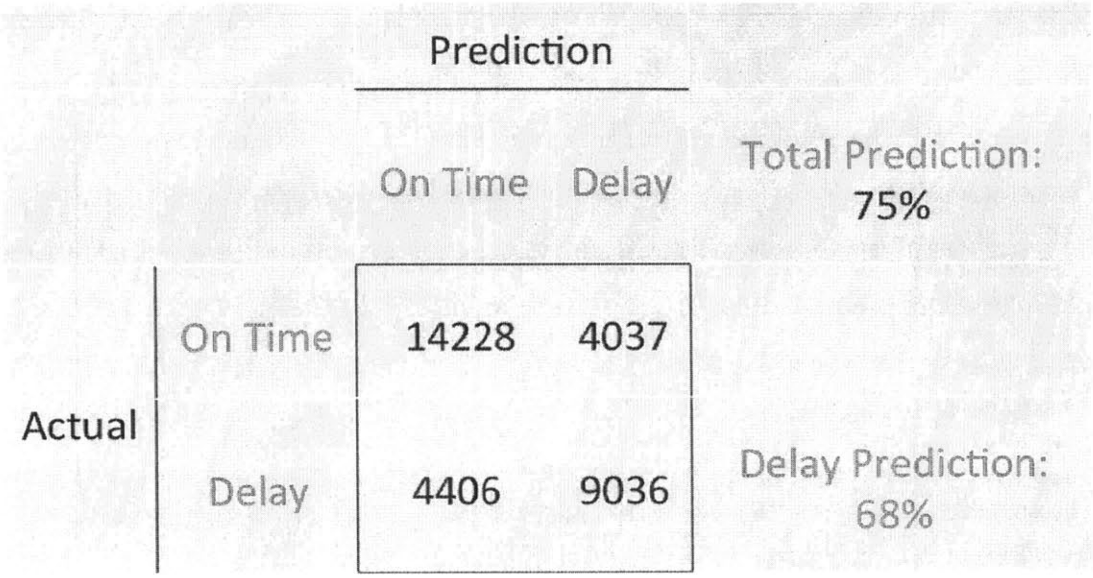
6.1 Random forest

Random forest is an ensemble learning method that combines multiple learning algorithms to obtain better predictive performance. Random forests algorithm based on the bagging method, and it generated n pruned trees using n bootstrap samples. In each tree, randomly sample mtry of the predictors and choose the best split among the subsets. For classification problem, the random forest growing an ensemble of trees and makes prediction based on the majority votes of trees (Breiman 2001).

Random forest performs well compared in prediction, and is able to handle large dataset efficiently. The learning algorithm is base on the Law of Large Numbers, and random forest does not have over fitting problem in general (Breiman 2001).

When processing invoice data, the first time customers have different payment pattern compared with returning customers ((Zeng et al. 2008). Therefore, we segregate two types of customer when building predictive model.

For returning customers who have at least two invoices in their account history, we use training data from November built prediction model using Random Forest algorithm. Then we test the performance of random forest model using training data. The confusion metrics of out of sample test is shows in Figure 6-1. The accuracy is 75% and the specificity is 68%.



		Prediction		
		On Time	Delay	Total Prediction: 75%
Actual	On Time	14228	4037	
	Delay	4406	9036	Delay Prediction: 68%

Figure 6-1 Confusion Matrix of November Invoice Prediction

The variable importance plot is an important output of random forest. It shows the relative importance of each variable in the model during the classification. The most importance variable showed in the top of the plot, while the least importance variable showed in the bottom. The importance plot of our delayed invoice predictive model shows in Figure 6-2.

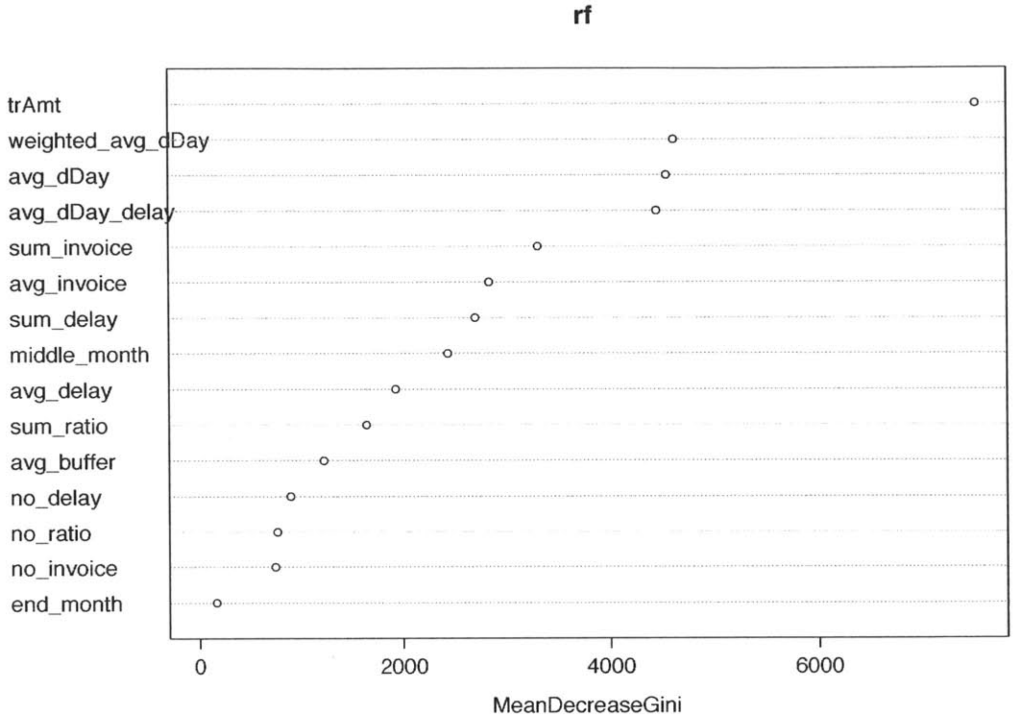


Figure 6-2 Variable Importance Plot

The importance plot shows the transaction amount on invoice level is the most importance variable in classifying delayed invoice. And the number of days of delay in customer level also very important, for example, the weighted average delay days, average delay days of invoices, and average delay days of delayed invoices for each customer.

Figure 6-3 shows the plot of classification error vs. number of trees grown in random forest model. From the plot, we can see the error become stable when n is close to 100. Therefore, in the model, we select number of trees equal to 100, which is stable and computational efficiency.

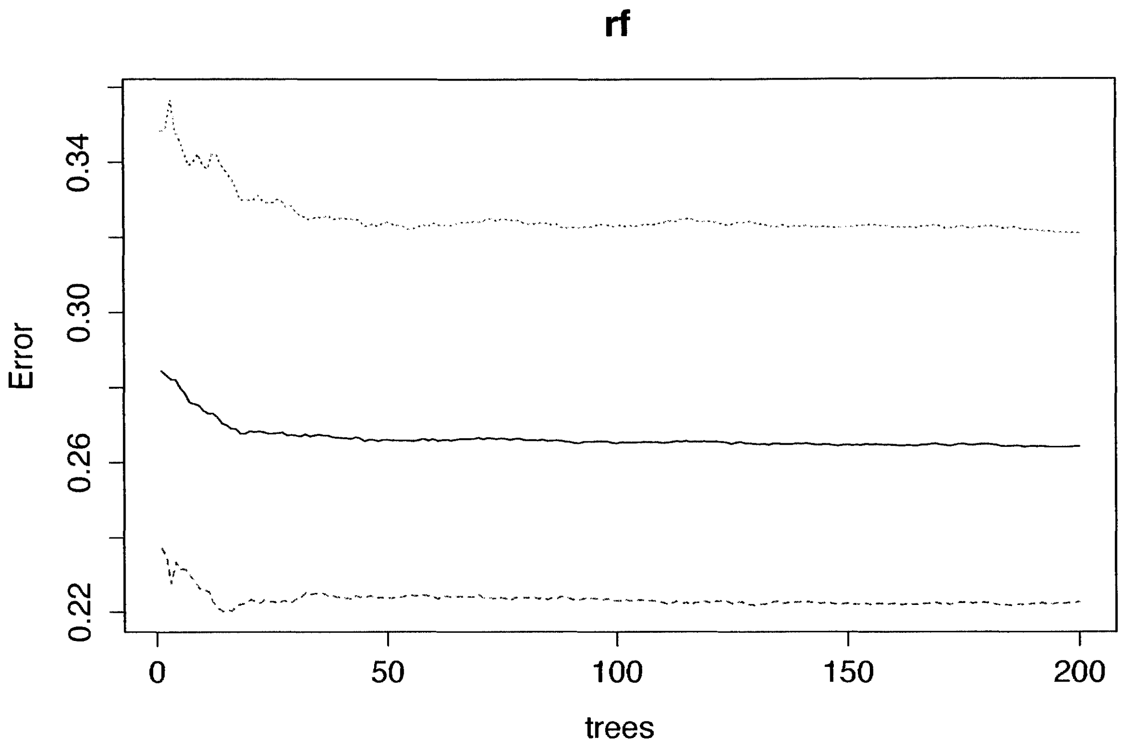


Figure 6-3 Classification Error vs Number of Trees Plot

6.2 Weighted Random Forest

The predictive model has relatively good performance in classification of invoices, and it has overall accuracy of 75%. For delay class, our model could capture 68% of delayed invoices. Then we rethink the methodology and comparable research work in invoice to cash, and figured out ways to improve the prediction accuracy of supervised learning model.

6.2.1 Group customer by number of invoices

In Table 3-2, we summarize the number of invoices and number of customers per month information. Using November data as an example, we divide the number of invoices by number of customer in this month and calculate the invoice per customer. We find the average number of invoices for each customer in November is about 1.5. Similar results also find in other months of data, and the average invoices per month are less than 2 for the rest of month's data. The detailed information of November invoices could be seen in Table 6-1.

Number of Invoices	Number of Customer	Invoices per Customer
409, 158	268, 622	1.52

Table 6-1 Invoice Summary of November

The composition of customers is an important factor of invoice delay prediction. In the invoice data from the technology firm, we find there are 135,566 first time customers in November, which is about 50% of total number customer in that month. This composition of customers is typical in the firm where the data comes from, and the proportion is consistent among other months. However, the distribution of customer behavior varies in different industries. In other related research work, researchers find the returning customers is more that 80% of their invoices (Zeng et al. 2008). The larger proportion of first time customers are not able to provide much historical information for the model, which affects the performance in the classification.

Another study mentioned that in their work, they find the average invoices number is about 15 invoices per customer per month (Hu 2015). The numerous monthly invoices provide sufficient historical delays information for customer and thus they were able to build a classification model with high prediction accuracy.

Therefore, I made a histogram to show the distribution of invoice per customer in Figure 6-4. From the histogram, we find the most of customers have only one or two invoices per month, which provide little historical information of the account. With no delayed information for the account, it is hard to make good prediction on the delayed behavior.

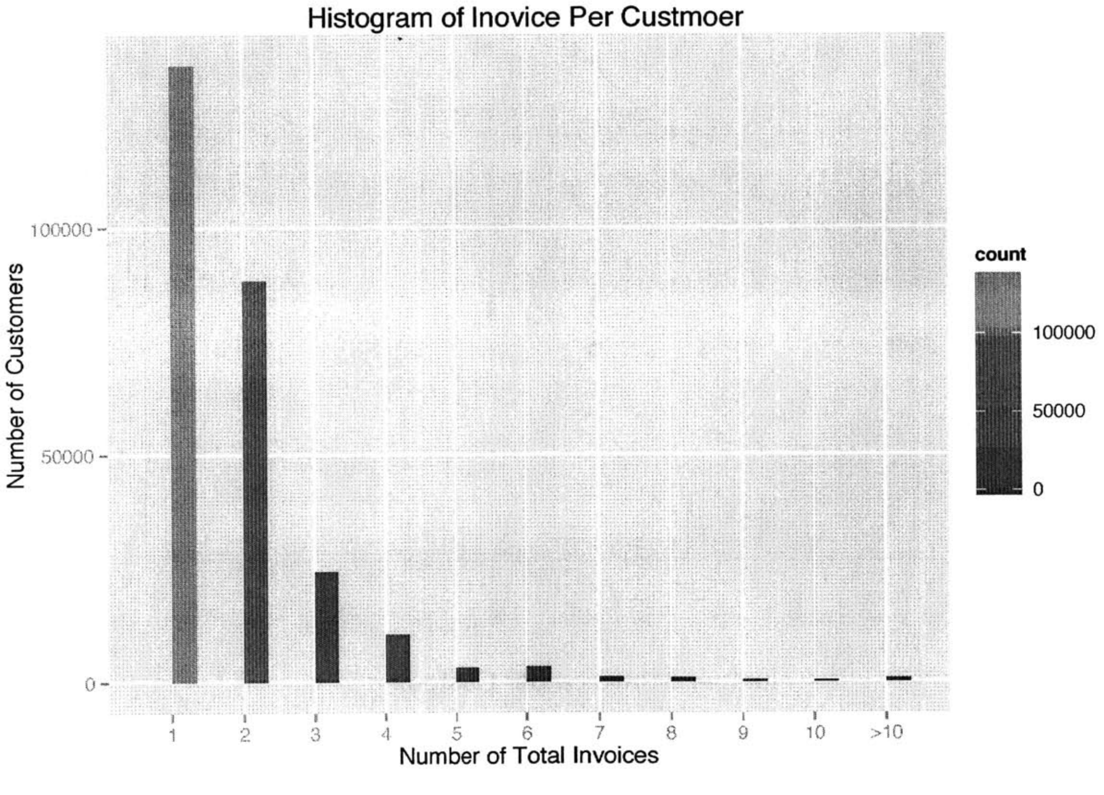


Figure 6-4 Histogram of Invoice Per Customer

In order to have a better understanding of customer's invoice history with their delayed behavior, I add a line on the top of histogram of the customer invoices summary in Figure 6-4. The line represents the percentage of delay for each customer group, which could be seen from Figure 6-5. From the plot, we find the customers with fewer invoices are less likely to have late payment on invoice, while customers with more invoices have a higher chance of delayed payment. Therefore, we find different customer group can have different payment behavior, which suggests us to build different models based on number of invoices for each customer group.

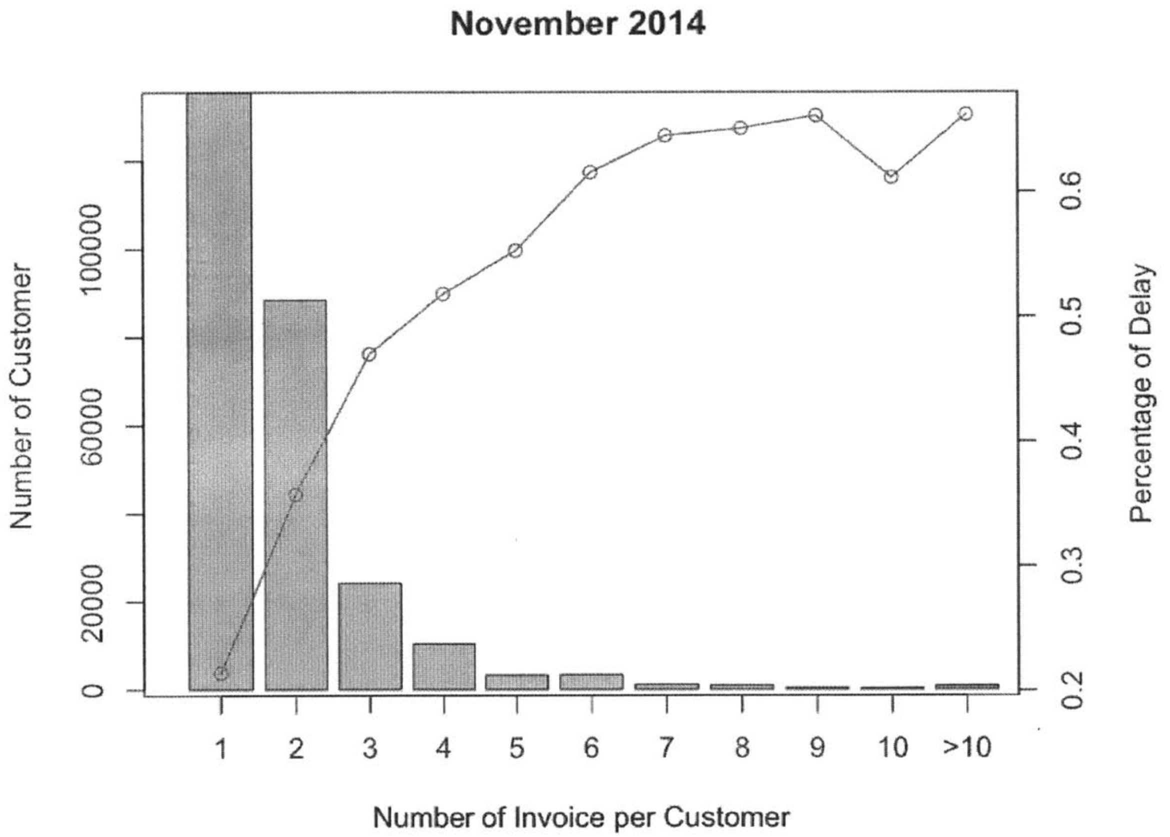


Figure 6-5 Customer Group and Delay Ratio

Table 6-2 shows the prediction results by different customer groups. Based on the analysis conclusion of Figure 6-5, we build ten machine-learning models for each customer group. For example, in case 3, we build the model using subset of data, which customer has at least 3 invoices. As seen from Table 6-2, the prediction accuracy increases as the number of invoices per customer increases. This results sounds logical for us, because the more invoices the customer has, the more historical information we know for the customer, which help us to have a better prediction on the customer's payment behavior. Therefore, when using the predictive model we have built before, it is necessary to check the number of invoices for customer, and make sure we have sufficient historical information for each account. We find the same results for the rest of months data.

Case	# of Invoices	Total Prediction	Delay Prediction
1	409,158	81.8%	55.6%
2	159,084	73.6%	67.5%
3	59,116	71.0%	66.9%
4	36,521	71.3%	71.3%
5	20,678	76.0%	77.8%
6	14,693	77.4%	79.9%
7	9,950	80.0%	82.5%
8	7,802	84.6%	85.1%
9	6,402	88.1%	90.8%
10	5,523	89.7%	94.3%

Table 6-2 Results by Customer Group

6.2.2 Weighted function in Random Forest

From data process section, we learned the invoice data set is imbalanced. The invoices data is composed by the on time invoice class with majority of the data points and the delayed invoice class with a few instances. Use November data as an example, in Table 4-4, there are only 28% of invoices belongs to delayed class, which shows the class distribution of invoices is highly skewed.

As a ensemble learning algorithm, random forest applies bootstrap samples, and induce classification trees by random features selection (Breiman 2001). While in handling imbalanced data, the random forest algorithm may be underrepresenting for minority class. The bootstrap sample may contain very few instances from minority class, which leads to a weak prediction performance in minority class (Chen 2004).

Cost sensitivity learning techniques should be added to the random forest to solve the imbalance data problem. In the invoice classification problem, the random forest classifier tends classify the invoice into the majority class, i.e. the on-time class. In the classification model, we are more interested in correctly identify the delayed invoices. Therefore, we need to add penalty on the misclassification of the minority class. In other words, we assigned weights to classes, and give larger weighted on the delayed class, which is the higher cost of misclassification. This approach of solving imbalance data based on random forest classifier is called weighted random forest (Chen 2004).

We investigated the weighted random forest with the November data, with weighted 1:2 and 1:3 for different customer groups. The results of the classification models showed in Table 6-3. The first column shows the overall accuracy, and the second column shows the specificity of random forest, weighted random forest (1:2), and weighted random forest (1:3). And we find the weighted random forest has superior performance compared to exist classifier.

Case	Accuracy	Specificity	Specificity (Weight 1:2)	Specificity (Weight 1:2)
1	74.4%	55.6%	89.6%	91.6%
2	70.2%	67.5%	85.2%	87.9%
3	70.1%	66.9%	77.8%	81.2%
4	71.5%	71.3%	81.0%	84.2%
5	76.0%	77.8%	83.9%	85.8%
6	77.7%	79.9%	86.6%	87.7%
7	80.8%	82.5%	89.2%	90.4%
8	84.8%	85.1%	91.1%	93.0%
9	88.4%	90.8%	94.8%	95.6%
10	89.0%	94.3%	96.8%	96.8%

Table 6-3 Weighted Random Forest Results

6.3 Robustness of the model

In the previous section, we present the prediction results of the modified random forest algorithm of delayed invoices. The classification model shows high accuracy in November invoice data. In this section, we were able to test the performance of the algorithm in two other months of invoice data.

In December and January, the sizes of data are comparable with November's invoices, which are about 100K per month. We went through same processes of machine learning for these new data. We used training sets of new invoice to train the model, and test the performance with testing sets.

The results of new invoices data also have high prediction accuracy, which shows relative consistent of the supervised learning algorithm in various data sets.

7 Conclusion

7.1 Summary

This research work has discussed the implementation of machine learning algorithms in the field of business analytics. The data sets we used through the project were the invoice data from a fortune 500 company in the technology market. This company issues invoices every month, and the size of invoices per month is around 100K. Of all the issued invoices from the company, only a small percentage of invoices have not been paid on time, which we defined as delayed invoices. The primary objective of the project is to forecast the delayed invoice in advance, using the historical delayed information, and build classification models using machine learning algorithm. Through the research work, we find that the supervised learning models are able to make significant improvement on the accuracy of delayed invoice prediction.

In the modeling part, I identified the pattern of data through statistical analysis, presented histogram and calculated the ratio of delayed invoice. Then I extracted significant features of invoice during data processing part. And I aggregated the data on customer level to provide additional information in perdition. The classification model is able to learn through selected features and provide accurate prediction results of delayed pattern of new issued invoices.

The invoice datasets from the firm have specific delay pattern on invoice collection. Most of invoices issued by the firm are paid on time, while small proportion of over due invoices, and the data has imbalanced class distribution. In the modeling part of work, I applied the weighted function in the machine learning algorithm and addressed the imbalance data issues. Another uniqueness of the datasets is the account history of customers. In the statistical analysis, I found about half of the accounts received the invoices have only one or two invoices in the history. The infrequent invoicing summary provides little historical information on the behavior of customers. Grouping customer by number of historical invoices is an approach that I implemented in the predictive modeling. Customer groups with more invoices in the history transaction have higher accuracy in prediction results. This approach increases the performance of classifier. The combination of two approaches could improve the accuracy of delayed invoice prediction.

7.2 Future work

Based on the framework of this study, there are several directions can be further studied.

- Incorporate with dispute information

In the study, we were given delay information of invoices, and we built supervised learning model with high prediction accuracy in delayed payment of invoice. These results can be further used in dispute invoice analysis given disputed information associated with invoices.

- Obtain more information in customer level

There is very little historical information for new customers in the data sets, and we could potentially extract other features that related to the payment behavior. One possible direction is to make predictions based on customers' profile data and get more detailed information from customers, i.e. industry, market capital, location, business type.

- Prioritize invoice collection

Based on the performance of predictive models, we can further explore the algorithm based on the invoice types and maximize business value. In other words, we could use our analysis results to optimize collection process.

8 Reference

1. Aleskerov, Emin, Bernd Freisleben, and Bharat Rao. 1997. “Cardwatch: A Neural Network Based Database Mining System for Credit Card Fraud Detection.” In *Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE 1997*, 220–26. IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=618940.
2. Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.
3. Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. “SMOTE: Synthetic Minority over-Sampling Technique.” *Journal of Artificial Intelligence Research*, 321–57.
4. Chen, C, A.Liaw, and L.Breiman. 2004. "Using random forest to learn imbalanced data, " Dept. of Statistics, U.C. Berkeley, Tech. Rep.
5. Correa, Alejandro, Andres Gonzalez, Catherine Nieto, and Darwin Amezquita. 2012. “Constructing a Credit Risk Scorecard Using Predictive Clusters.” In *SAS Global Forum*, 128. http://albahnsen.com/files/Constructing%20a%20Credit%20Risk%20Scorecard%20using%20Predictive%20Clusters.pdf.pdf.
6. Dorronsoro, J.R., F. Ginel, C. Sgnchez, and C.S. Cruz. 1997. “Neural Fraud Detection in Credit Card Operations.” *IEEE Transactions on Neural Networks* 8 (4): 827–34. doi:10.1109/72.595879.
7. Drosou, Krystallenia, Stelios Georgiou, Christos Koukouvinos, and Stella Stylianou. 2014. “Support Vector Machines Classification on Class Imbalanced Data: A Case Study with Real Medical Data.” *Journal of Data Science* 12 (4): 143–55.
8. Erdmann, Tashi P., Manon de Groot, and Ronald J. M. M. Does. 2010. “Quality Quandaries: Improving the Invoicing Process of a Consulting Company.” *Quality Engineering* 22 (3): 214–21. doi:10.1080/08982111003771854.

9. Fawcett, Tom. 2006. “An Introduction to ROC Analysis.” *Pattern Recognition Letters* 27 (8): 861–74. doi:10.1016/j.patrec.2005.10.010.

10. He, Hongxing, Warwick Graco, and Xin Yao. 1998. “Application of Genetic Algorithm and K-Nearest Neighbour Method in Medical Fraud Detection.” In *Simulated Evolution and Learning,* 74–81. Springer. http://link.springer.com/chapter/10.1007/3-540-48873-1_11.

11. Hu, Peiguang. 2015. “Predicting and Improving Invoice-to-Cash Collection through Machine Learning.” Massachusetts Institute of Technology. http://dspace.mit.edu/handle/1721.1/99584.

12. "Introduction to ROC Curves." Introduction to ROC Curves. Accessed May 18, 2016. http://gim.unmc.edu/dxtests/roc1.htm.

13. Khandani, Amir E., Adlar J. Kim, and Andrew W. Lo. 2010. “Consumer Credit-Risk Models via Machine-Learning Algorithms.” *Journal of Banking & Finance* 34 (11): 2767–87.

14. Kokkinaki, Angelika I. 1997. “On Atypical Database Transactions: Identification of Probable Frauds Using Machine Learning for User Profiling.” In *Knowledge and Data Engineering Exchange Workshop, 1997. Proceedings,* 107–13. IEEE. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=629848.

15. Ling, Charles X., and ghui Li. 1998. “Data Mining for Direct Marketing: Problems and Solutions.” In *KDD,* 98:73–79. http://www.csd.uwo.ca/~cling/papers/kdd98.pdf.

16. Maimon, Oded, and Lior Rokach. 2005. *Decomposition Methodology for Knowledge Discovery and Data Mining.* Springer. http://link.springer.com/chapter/10.1007/0-387-25465-X_46.

17. Menardi, Giovanna, and Nicola Torelli. 2014. “Training and Assessing Classification Rules with Imbalanced Data.” *Data Mining and Knowledge Discovery* 28 (1): 92–122. doi:10.1007/s10618-012-0295-5.

18. Ormerod, Thomas, Nicola Morley, Linden Ball, Charles Langley, and Clive Spenser. 2003. “Using Ethnography to Design a Mass Detection Tool (MDT) for the Early Discovery of Insurance Fraud.” In *CHI’03 Extended Abstracts on*

Human Factors in Computing Systems, 650–51. ACM. http://dl.acm.org/citation.cfm?id=765910.

19. Phua, Clifton, Vincent Lee, Kate Smith, and Ross Gayler. 2010. "A Comprehensive Survey of Data Mining-Based Fraud Detection Research." *arXiv Preprint arXiv:1009.6119*. http://arxiv.org/abs/1009.6119.
20. "ROC Curves." MedCalc. Accessed May 18, 2016. https://www.medcalc.org/manual/roc-curves.php.
21. "ROC Curve Demonstration." ROC Curve Demonstration. Accessed May 18, 2016. http://arogozhnikov.github.io/2015/10/05/roc-curve.html.
22. Simić, Dragan, Svetlana Simić, and Vasa Svirčević. 2011. "Invoicing and Financial Forecasting of Time and Amount of Corresponding Cash Inflow." *Management Information Systems* 6 (3): 14–21.
23. Sutton, Oliver. 2012. "Introduction to K Nearest Neighbour Classification and Condensed Nearest Neighbour Data Reduction." *University Lectures, University of Leicester*. http://www.math.le.ac.uk/people/ag153/homepage/KNN/OliverKNN_Talk.pdf.
24. Yamanishi, Kenji, Jun-Ichi Takeuchi, Graham Williams, and Peter Milne. 2004. "On-Line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms." *Data Mining and Knowledge Discovery* 8 (3): 275–300.
25. Younes, Bashar, Ahmed Bouferguène, Mohamed Al-Hussein, and Haitao Yu. 2015. "Overdue Invoice Management: Markov Chain Approach." *Journal of Construction Engineering and Management* 141 (1): 04014062. doi:10.1061/(ASCE)CO.1943-7862.0000913.
26. Zeng, Sai, Prem Melville, Christian A. Lang, Ioana Boier-Martin, and Conrad Murphy. 2008. "Using Predictive Analysis to Improve Invoice-to-Cash Collection." In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1043–50. ACM. http://dl.acm.org/citation.cfm?id=1402014.

27. Zweig, Mark H., and Gregory Campbell. 1993. "Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine." *Clinical Chemistry* 39 (4): 561–77.