A Novel Approach for Forecasting Account Receivables

by Bintu Kadiwala

Submission date: 30-May-2021 05:12PM (UTC+0530)
Submission ID: 1597030465
File name: A_Novel_Approach_for_Forecasting_Account_Receivables.pdf (229.11K)
Word count: 3797
Character count: 19621

A Novel Approach for Forecasting Account Receivables

Parth Kapadia[1], Bintu Kadhiwala[2], Tejaswini Bahurupi[3], Het Dalal[4], Siddhi Jariwala[5], and Kshitij Naik[6]

Computer Engineering Department, Sarvajanik College of Engineering and Technology, Surat, India

kapadia.parth1999@gmail.com[1], bintukadhiwala@gmail.com[2], tejaswinibahurupi@gmail.com[3], hetjd95@gmail.com[4], skjjariwala@gmail.com[5], kshitijnaiK2017@gmail.com[6]

Abstract. In recent times, various firms/companies provide services to their customers on credit. These firms receive the payments, against the services and/or purchases provided to the customers but not paid for, termed as Account Receivables. With an aim to plan the finance in addition to decide strategies for their business, the executives of these firms desire to predict the Account Receivables of their companies for a specific timespan. For the said purpose, in this paper, we present three different methods. These methods utilize Machine Learning prediction models - Logistic Regression, Random Forest classifier, and an ensemble of K-Means clustering and Random Forest classifier discussed in the literature. However, with an aim to find the optimal solution in terms of prediction accuracy, we put forward the theoretical and the experimental analysis of these methods. Experimental results affirm that an ensemble of K-Means clustering and Random Forest classifier outperforms among three methods.

Keywords: Account Receivables, Machine Learning, ensemble, forecasting

1 Introduction

In the corporate world, most of the customers purchase the goods/services on credit and promise to pay the amount to the firm/company on a particular scheduled date or within a specific timespan (Fig. 1). For the services and/or purchases provided to the customers but not paid for, the firms receive the payments from the customers popularly termed as 'Account Receivables'. Usually, the companies give their customers a limit of 30, 60, or 90 days to pay, and record this transaction as an invoice. The company lists this amount to be paid by the customer in its Account Receivables. Once the company receives the payment for the particular invoice, the amount is inserted into the sales account of the business.

Thus, Account Receivables is the money or the payment a customer owes to the company for the purchases they make on credit. On the balance sheet of

the company, it is considered to be an asset for the company as it signifies the money that is to be received by the company. Therefore, Account Receivables is an important aspect for calculating the profits of the company. The executives of many companies allow a certain percentage of their sales to be on credit with an aim to increase their sales and to reduce the transaction cost [1]. As discuss in [2], Account Receivables exceed a quarter of the total corporate assets in countries like France, Germany, and Italy. Moreover, the authors of [3] find that 18% of the total assets of US firms consist of receivables. As per our observations, the company must collect the outstanding amount within a certain time period to gain maximum benefits from Account Receivables.

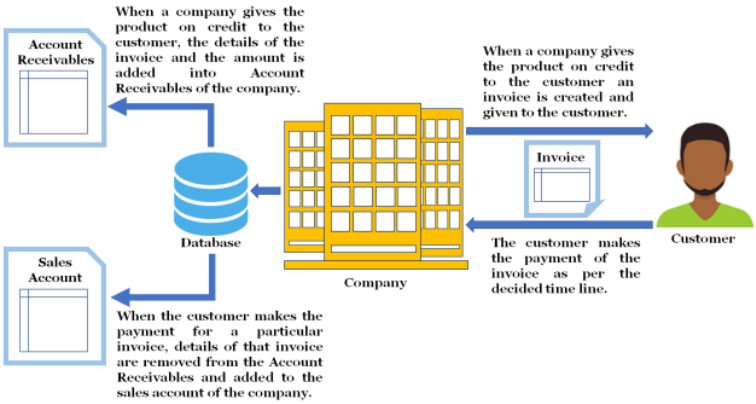

Fig. 1. Working of Account Receivables

The average number of days a company takes to collect Account Receivables amount from their customers is termed as 'Days Sales Outstanding' that gives an idea of how quickly a company can collect its outstanding money. The greater value for Days Sales Outstanding indicates that the company takes more time to collect money that can lead to cash flow problems. On the other hand, the lower value for the same indicates that the company collects the money in less time. In addition, the effectiveness of the company to convert the Account Receivables into cash is measured through the 'Receivables Turnover Ratio' that is a comparison between credit sales and average receivables [4]. It indicates how efficiently a company can collect the amount against the services provided to the customers on credit. Whenever the company fails to recollect this amount, it turns out to be a bad debt resulting in a loss for the company. Hence, with an aim to avoid such losses, it is highly required by the executives of the companies (1) to manage the Account Receivables, and (2) to forecast the Account Receivables expected to receive in a particular time span.

Consequently, forecasting of the Account Receivables plays a major role for the company executives as it can have a large impact on the working capital of the company. By forecasting, the company gets an insight into the cash amount to be received from the customers against services over a certain time period. For this purpose, we train a model that predicts the probability of the customer paying the particular amount within 30, 60, and 90 days after the due date. Hence, the key objective of this paper is to find the optimal solution with the maximum prediction accuracy for forecasting Account Receivables. This forecasting definitely helps the company executives plan the finances of the company efficiently.

The rest of the paper is organized as follows: Section-II discusses the related work. Section-III describes various machine learning methods that we utilize for the purpose of the prediction and the real-time transaction dataset on which the experiments are performed. The experimental results obtained are debated and analysed at length in Section-IV. Finally, Section-V puts concluding remarks on our work.

2 RELATED WORK

In [5], the authors work on the prototype that is able to support collectors in predicting the payment of invoices. They focus on the problems faced by the collectors to contact the customers on the payment due date for collection of Account Receivables. For the same, they provide the solution through a predictive modelling approach to prioritize contacting clients based on the probability of late payment helps the collector to make a more effective and efficient decision. The major objective of their study is to make an assertive and timely decision by focusing on prioritizing the collection of invoices from the customers. This is decided based upon, (1) the possibility of higher financial return, and (2) the customers that are most likely to make the payment. They utilize different machine learning models such as Naive Bayes, Logistic Regression, K-Nearest Neighbors, Gradient Boost Decision Tree, Deep Neural Network. Their model is limited to prioritize the invoice payment and cannot predict when a particular payment will be made.

In [6], with the help of historical data, the author demonstrates how machine learning models predict payment outcomes of invoices that are not paid. Here, the author divides the days past due date into six buckets viz. 1-30 days, 31-60 days, and so on, in which the invoices can be paid and tries to predict the bucket in which the Account Receivables can be received. For the same, a predictive analysis approach is used to quantify the effect of the future decision in order to advise on possible outcomes before making the actual decision. Here, the author uses the historical data of the customers paid invoices and tries to identify the payment behaviour of customers to create a predictive model. This model predicts the payment bucket of the newly credited invoice with the help of the AdaBoost Decision Tree algorithm. The accuracy attained by this approach is better than the approach discussed in [4]. However, the accuracy decreases gradually from bucket 1 to bucket 6 that ultimately results in only 50% accuracy for

bucket 6. As described by the author in [6], the model is unable to handle those invoices which are overdue and cross the dispute limit.

In [7], the authors demonstrate how supervised learning is used to build predictive models for the payment outcomes prediction of newly created invoices. Their model predicts whether an invoice is paid on time or not and also provides the delay estimation. Here, the task of prediction is formulated with the help of a model that classifies the new instance(s)/invoice(s) into five target classes. The authors study various algorithms namely C4.5 Decision Tree Induction, Naive Bayes, and PART algorithm, and conclude that the PART algorithm gives the best performance in terms of classification accuracy. For the same, they develop a set of aggregated features that capture historical payment behaviour for each customer to make the prediction. Through this model, the company gets only a probable idea of whether the Account Receivables is paid by the customer or not. However, this model cannot predict the exact timespan for paying the Account Receivables.

3 THEORETICAL BACKGROUND

3.1 Machine Learning Methods Considered

In this subsection, we summarize three methods that we utilize for the purpose of the classification.

3.1.1 Logistic Regression : The examination of dichotomous or parallel classes is done with the help of a Logistic Regression model. Although it provides us with the capacity to adjust various classes, in our case we utilize the same to handle only two classes [8]. This makes strategic relapse particularly valuable for the investigation of observational information when change is expected to diminish. The normal method for fitting the Logistic Regression model has optimal properties [9], however, we observe while performing experiments that it is extremely sensitive to noisy data. For experimentations, we build a Logistic regression model in python using "sklearn.linear_model.LogisticRegression" package in sci-kit learn module.

3.1.2 Random Forest Classifier : The Random Forest classifier is an ensemble of several decision tree models for a set of data. For each tree, the data is split into units which are commonly referred as nodes. Split focus depends on estimations of indicator factors. Subsequently, factors used to part the information is viewed as significant illustrative factors. Random forest fits these separate decision trees to bring out a collective decision [10]. For experimentations and to have uniformity in findings, we also build the Random Forest classifier model in python using "sklearn.ensemble.RandomForestClassifier" package in sci-kit learn module and we keep the ensemble of 100 decision trees.

3.1.3 Ensemble of K-Means Clustering and Random Forest Classifier : The use of ensemble classifiers is escalated recently as the main idea of the ensemble is to combine a set of models, each of which performs its own typical task with an aim to gain a composite model with improved accuracy. Alternatively, an ensemble classifier can also be viewed as a technique of combining many weak

learners to produce a strong learner [11]. Additionally, in the past few years, experimental studies show that combining the outputs of multiple classifiers i.e., ensemble methods reduce the generalization error [12]. In our work, we utilize an ensemble of the K-Means clustering and Random Forest classification algorithm as Random Forest classifier is not able to remove redundancies [13] that can be removed with the help of a clustering algorithm. As pre-processing is a process of removing redundant features, we employ a clustering approach using the K-Means algorithm to group similar features [14]. For experimentations, we preprocess the data using "sklearn.cluster.KMeans" package and build the Random Forest classifier model in python using "sklearn.ensemble.RandomForestClassifier" package in sci-kit learn module and we also keep the ensemble of 100 decision trees.

3.2 Transaction Dataset Description

We perform experiments on the live dataset provided by Sailfin Technologies Pvt. Ltd. [15]. The dataset consists of 51,53,556 closed transactions of one of the company's customers. The dataset is comprising of the following six parameters:

1. ID - Transaction ID
2. Account Number - Account from which transaction takes place
3. Amount - Amount of the Account Receivables
4. Create Date - Date on which the Account Receivables is created
5. Due Date - Due date for closing the Account Receivables
6. Close Date - Date on which the Account Receivables is closed, i.e. full payment is made

3.3 Extracted Parameters

For experimentation, we build a classification model with the help of four extracted parameters. In this subsection, we summarize these parameters extracted from the transaction dataset (subsection-3.2) along with the extraction methods.

1. Parameter - 1 (Normalized Amount) and Parameter - 2 (Normalized DPD): Using normalizing techniques over the datasets, the Normalised Amount and Normalized DPD are calculated as follows:
   - For method 1 (subsection 3.1.1) and method 2 (subsection 3.1.2):
     (a) Normalized Amount = Actual amount/Average amount of dataset
     (b) Normalized DPD = Actual DPD/Average DPD of dataset
   - For method 3 (subsection 3.1.3):
     (a) Normalized Amount = Actual amount/Average amount of cluster
     (b) Normalized DPD = Actual DPD/Average DPD of cluster

   Where, Days Past Due (DPD) is calculated using, DPD = (Due date - Close date)
2. Parameter - 3 (EPID) and Parameter - 4 (EPDD):
   It is observed from the dataset that all payments are made by the customers either in the month of due date or later than that. Hence, we consider two parameters - EPID and EPDD as follows:
   (a) EPID = (End of month of invoice due date - Invoice create date)
   (b) EPDD = (End of month of invoice due date - Due date)

4 EXPERIMENTAL ANALYSIS

4.1 Experimental Setup

In this paper, the forecasting of Account Receivables for a specific month is treated as a two-class classification problem. One class out of these two classes handles those Account Receivables that is credited in a specific month whereas another handles those Account Receivables that is not credited. Hence, it is considered as a binary classification problem with two class labels - 'yes(1)' and 'no(0)' for the Account Receivables either credited or not credited respectively in a specific timespan (month).

Table 1. Data distribution for classification in current month (80:20 split)

	Collected in current month (Class label: 1)	**Not collected in current month (Class label: 0)**	**Total**
Training data	16,93,964	24,28,880	41,22,844
Testing data	4,23,368	6,07,344	10,30,712
Total	21,17,332	30,36,224	51,53,556

Table 2. Data distribution for classification in next month (80:20 split)

	Collected in next month (Class label: 1)	**Not collected in next month (Class label: 0)**	**Total**
Training data	22,63,066	18,59,778	41,22,844
Testing data	5,65,392	4,65,320	10,30,712
Total	28,28,458	23,25,098	51,53,556

Table 3. Data distribution for classification in next to next (n2n) month (80:20 split)

	Collected in n2n month (Class label: 1)	**Not collected in n2n month (Class label: 0)**	**Total**
Training data	24,45,042	16,77,802	41,22,844
Testing data	6,10,743	4,19,969	10,30,712
Total	30,55,785	20,97,771	51,53,556

From the transaction dataset and the general corporate finance [16], we observe that the due invoices are closed within a maximum of 90 days after the due date. Hence, we train three different classification models with an aim to determine whether the invoices will be collected in the month of the due date (current month), in the next month of the due date (next month), or in the next-to-next month of the due date (n2n month). Thus, we calculate the output parameter as - whether the invoices are collected in the current month or not (1 or 0 respectively), the next month or not (1 or 0 respectively), and the next-to-next month or not (1 or 0 respectively) for three models respectively.

Furthermore, for all the three cases - if an invoice is predicted to be closed in the current month of the due date, it is certain that it will be predicted as closed in the next month of the due date too. Similarly, if an invoice is predicted to be closed in the next month of the due date, then it will be predicted as closed in the next-to-next month of the due date too.

For experimental analysis, we perform experiments on aforementioned Transaction Dataset (section 3.2) for various Train Data to Test Data split - 60:40, 70:30, 80:20, and 90:10. To have a clear insight, we describe the distribution of the data for 80:20 Train Data to Test Data split for current month (Table 1), next month (Table 2) and n2n month (Table 3).

Furthermore, the transaction data having a negative amount and/or high DPD is considered as noisy data and hence, it is highly required to remove this data. For the purpose, we use the Pandas library of Python.

4.2 Experimental Results

The key findings from the obtained results (Fig. 2, 3, and 4) are as follows:

- From Fig. 2, 3, and 4, it is easily observed that the Logistic Regression model gives relatively less accuracy among all the three models. The major limitation of this model is that it keeps linear relation between all the dependent and independent variables. The classification problem addressed in this paper is clearly a non-linear problem and hence, the accuracy resulting from Logistic Regression model is not acceptable.
- From Fig. 3, it is seen that Random Forest classifier outperforms the Logistic Regression model. As mentioned (section 3.1.3), Random Forest classifier fails to remove redundancies [12]. Hence, the probability of selecting an irrelevant feature is high because of its random nature of selection (from 100 decision trees taken into account as discussed earlier). This is the reason why a comparatively low accuracy is observed for some categories (Fig. 3).
- From Fig. 4, it is observed that the ensemble of K-Means clustering and Random Forest classifier outperforms previously considered two models in all aspects. Here, the key drawback of Random Forest classifier for not being able to remove redundant data is eliminated by adding K-Means clustering. The Amount and DPD are first passed as parameters in the clustering algorithm as they are the main factors for redundancy that requires to be eliminated. Hence, we observe that the ensemble of Random Forest classifier and K-Means clustering results into the highest accuracy amongst three classification models.

5 CONCLUSION

In this paper, we utilize three Machine Learning methods to forecast the Account Receivables that any company would receive in a specific time period. We build three models using each of these methods to predict the Account Receivables being closed in (1) the month of the due date, (2) current or next month of the due date, and (3) current or next or next-to-next month of the due date.

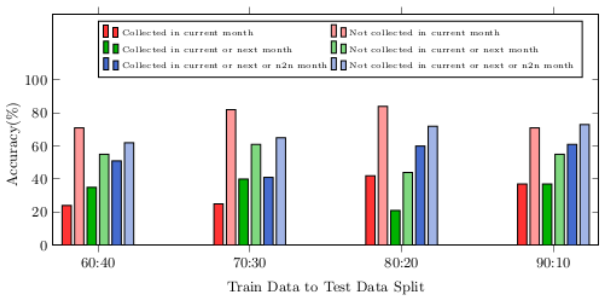

Fig. 2. Accuracy Vs. Train Data to Test Data Split (Logistic Regression model)

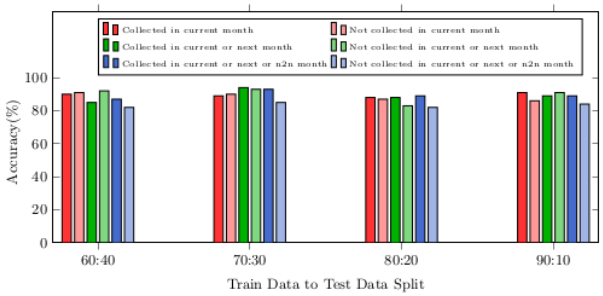

Fig. 3. Accuracy Vs. Train Data to Test Data Split (Random Forest classifier)

Fig. 4. Accuracy Vs. Train Data to Test Data Split (Ensemble of K-Means clustering and Random Forest classifier)

Additionally, we analyze the results obtained through these models for different Train Data to Test Data split. From this analysis, we conclude that an ensemble of Random Forest classifier and K-Means clustering outperforms among three methods. Hence, the findings of this paper assist the business executives to forecast the Account Receivables in advance with an aim to gain better idea of finances and decide strategies for their varied businesses.

References

1. Abuhommous, A.A., Mashoka, T.: A dynamic approach to accounts receivable: the case of jordanian firms. Eurasian Business Review 8(2), 171–191 (2018), doi: https://doi.org/10.1007/s40821-017-0074-8
2. Maksimovic, V.: Firms as financial intermediaries: Evidence from trade credit data. The World Bank (2001), doi: https://doi.org/10.1596/1813-9450-2696
3. Rajan, R., Zingales, L.: Financial dependence and growth, american economic review (1998)
4. Purwanti, T.: An analysis of cash and receivables turnover effect towards company profitability. International Journal of Seocology pp. 037–044 (2019), doi: https://doi.org/10.29040/seocology.v1i01.6
5. Appel, A.P., Malfatti, G.L., Cunha, R.L.d.F., Lima, B., de Paula, R.: Predicting account receivables with machine learning. arXiv preprint arXiv:2008.07363 (2020)
6. Shah, H.: Customer payment prediction in account receivable. International Journal of Science and Research (IJSR) 8(1), 642–644 (2019)
7. Zeng, S., Boier-Martin, I., Melville, P., Murphy, C., Lang, C.A.: Predictive modeling for collections of accounts receivable. In: Proceedings of the 2007 international workshop on Domain driven data mining - DDDM '07. pp. 43–48 (2007), doi: https://doi.org/10.1145/1288552.1288558
8. Valley, M.P.: Logistic regression. Circulation 117(18), 2395–2399 (2008), doi: https://doi.org/10.1161/CIRCULATIONAHA.106.682658
9. Pregibon, D., et al.: Logistic regression diagnostics. The Annals of statistics 9(4), 705–724 (1981), doi: https://doi.org/10.1214/aos/1176345513
10. Everingham, Y., Sexton, J., Skocaj, D., Inman-Bamber, G.: Accurate prediction of sugarcane yield using a random forest algorithm. Agronomy for sustainable development 36(2), 27 (2016), doi: https://doi.org/10.1007/s13593-016-0364-z
11. Pappu, V., Pardalos, P.M.: High-dimensional data classification. In: Clusters, orders, and trees: methods and applications, pp. 119–150. Springer (2014), doi: https://doi.org/10.1007/978-1-4939-0742-7_8
12. Ditterich, T.: Ensemble methods in machine learning, multiple classifier systems (2000)
13. Darbon, J., Osher, S.: Algorithms for overcoming the curse of dimensionality for certain hamilton–jacobi equations arising in control theory and elsewhere. Research in the Mathematical Sciences 3(1), 1–26 (2016), doi: https://doi.org/10.1186/s40687-016-0068-7
14. Aydadenta, H., Adiwijaya, A.: A clustering approach for feature selection in microarray data classification using random forest. Journal of Information Processing Systems 14(5), 1167–1175 (2018)
15. https://www.sailfin.tech/
16. Bloomenthal, A.: How long can accounts receivables remain outstanding? (2020), https://www.investopedia.com/ask/answers/021215/how-long-are-accounts-receivable-allowed-be-outstanding.asp

A Novel Approach for Forecasting Account Receivables

ORIGINALITY REPORT

11%	10%	8%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	export.arxiv.org Internet Source	1%
2	link.springer.com Internet Source	1%
3	Stephen Leo, Massimiliano De Antoni Migliorati, Peter R. Grace. "Predicting within - field cotton yields using publicly available datasets and machine learning", Agronomy Journal, 2020 Publication	1%
4	Ramendra Pratap Singh, Ramendra Singh, Prashant Mishra. "Does managing customer accounts receivable impact customer relationships, and sales performance? An empirical investigation", Journal of Retailing and Consumer Services, 2021 Publication	1%
5	seocologi.com Internet Source	1%

6	S Zainuddin, F Nhita, U N Wisesty. "Classification of gene expressions of lung cancer and colon tumor using Adaptive-Network-Based Fuzzy Inference System (ANFIS) with Ant Colony Optimization (ACO) as the feature selection", Journal of Physics: Conference Series, 2019 Publication	1 %
7	cesmaa.org Internet Source	1 %
8	www.investopedia.com Internet Source	1 %
9	www.tandfonline.com Internet Source	1 %
10	Shadi Abpeykar, Mehdi Ghatee. "Neural trees with peer-to-peer and server-to-client knowledge transferring models for high-dimensional data classification", Expert Systems with Applications, 2019 Publication	1 %
11	"Data Analytics on Agrometeorological Parameters for Building A Utility System for Farmer Community", International Journal of Recent Technology and Engineering, 2019 Publication	<1 %
12	ns2.thinkmind.org Internet Source	<1 %

13	dyuthi.cusat.ac.in Internet Source	<1 %
14	Poojan Dalal, Parth Kapadia, Bhavik Jinjala, Dhatri Pandya. "Autonomous Inpainting Algorithm for Wire Removal from Image", 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), 2021 Publication	<1 %
15	dokumen.pub Internet Source	<1 %
16	www.springerprofessional.de Internet Source	<1 %
17	Godson K. Mensah, Werner D. Gottwald. "Enterprise risk management: Factors associated with effective implementation", Risk Governance and Control: Financial Markets and Institutions, 2016 Publication	<1 %
18	academic.oup.com Internet Source	<1 %
19	www.ijrte.org Internet Source	<1 %
20	Mahesh L. Maskey, Tapan B Pathak, Surendra K. Dara. "Weather Based Strawberry Yield Forecasts at Field Scale Using Statistical and Machine Learning Models", Atmosphere, 2019 Publication	<1 %

21	circ.ahajournals.org Internet Source	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	Off		