International Journal of Research in Engineering, Science and Management
Volume 6, Issue 11, November 2023
https://www.ijresm.com | ISSN (Online): 2581-5792

Machine Learning for Accounts Receivable Payment Forecasting

M. Raj Mani*
Research Scholar, Department of Computer Science Engineering, CMR Institute of Technology, Bengaluru, India

***Abstract*: Account receivables (AR) are a business's most valuable asset. This research paper explores how supervised machine learning can predict payment outcomes for open invoices, specifically addressing the common challenge of maintaining consistent income for small to medium enterprises (SMEs). We developed a model using various machine learning techniques, including linear regression, random forest regressor, decision tree classifier, Linear SVR, and XGB Regressor, by training it on actual AR data. This model assists collectors in forecasting debt payment.**

***Keywords*: Logistic regression, XGBoost, Decision Tree, Random Forest, Payment prediction.**

1. Introduction

In the field of business analytics (BA), the big data age has begun. Walmart, for instance, processes more than 1 million transactions every hour and maintains databases containing over 2.5 petabytes (2.5 * 101) of data. Estimates indicate that by 2009, the US economy's roughly all sectors had at least 200 terabytes (2 * 1012) of data on average [1].

Business analytics is the discipline of methodically and iteratively examining data from an organization with a focus on statistical analysis. Businesses that employ a structured data collection process use it to decide based on data. However, the business analytics has undergone a tremendous transition due to a significant increase in data available during the previous 15 years [2]. In particular, this data flood necessitates automated data analysis techniques, where academics turned to machines for assistance.

Predictive models are developed using past customer data, customer event patterns, or collector's notes for a customer. Using neural networks as a predictive modeling tool is one approach. Delays and numbers are equally important for collection decision-making, despite the approach's concentration on anticipating collection quantities. We also can handle collections at the client and invoice levels [3].

One's ability to convert data into knowledge that may be used is strengthened via machine learning. Five methods for using big data that are broadly applicable and can change sectors have been discovered by a business, McKinsey & Co. [4].

- Transparency in business information
- Facilitating experimentation to identify needs, reveal variability, and enhance performance.
- Dividing up populations to create tailored responses.
- Automating or assisting human decision-making via algorithms.
- Developing fresh products, services, and business concepts.

2. Methodology

The goal of predicting future invoice payments falls within the category of supervised machine learning since it relies on past customer payment information. We employed the XGBoost, Decision Tree Classifier, Random Forest Regressor, Linear Regression, Linear SVR, and XGB Regressor method. These issues are classified as multiclass because there are five potential solutions.

These age buckets represent the time before or after the due date that an invoice is paid. Therefore, if an invoice's due date is January 15 and the payment is made on February 25, then the age bucket for that invoice is "16- 30."

Table 1
Age bucket

S.No.	Aging Bucket	Value Counts
1	0-15	7735
2	16-30	169
3	31-45	69
4	46-60	6
5	Greater than 60	5

A. XGBoost

Gradient XGBoost technology is applied to develop boosted decision trees. Decision trees are created by using the approach sequentially. Weights play a significant role with XGBoost. Each independent variable is weighed before being added to the decision tree that forecasts outcomes. Before entering the second decision tree, additional weights given to the variables were mistakenly predicted by the first model. Then, a robust and accurate model is created by combining these various classifiers and predictors. Regression, classification, ranking, and custom prediction problems can all be resolved with it.

The following definition summarizes the above model's function:

$$\hat{y}_i^{(0)} = 0 \qquad (1)$$

$$\hat{y}_i^{(1)} = f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \qquad (2)$$

*Corresponding author: rajveersinha694@gmail.com

$$\hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \quad (3)$$
$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} + f_1(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (4)$$

B. Decision Tree

In a decision tree, every internal node represents a test on an attribute, every branch is a test result, and every leaf node (terminal node) defines as a class level. It is a type of tree structure similar to a flow table.

Entropy: Entropy is a measure of a random variable's uncertainty, and it identifies the impurity in any given collection.

$$Gain(S, A) = Entropy(S) \sum Values(A) \frac{S_v}{S} \; Entropy\ (S_v) \quad (5)$$

C. Linear Regression

It performs a regression operation. Regression takes independent variables to model a goal prediction value. It does so to establish a relationship between the variables and forecasting.

Linear Regression Hypothesis Function:

$$y = \theta + \theta_2 . x \quad (6)$$

D. Linear SVR

A method to resolve regression issues in the linear SVR algorithm. The linear SVR algorithm, which employs the linear kernel strategy, successfully handles large datasets.

E. Random Forest

For predicting the accuracy of the input dataset, the Random Forest classifier takes the averages of all the results from many decision trees that are applied to several subsets of the input dataset.

This approach is broken down into six steps, similar to any supervised machine learning procedure.

1. *Data Pre-Processing:* In this process, we are removing the constant column and quasi-constant column
2. *Data Splitting:* We are splitting the data into 3 dataset
   i) Training
   ii) Validation
   iii) Testing
3. *Data Visualization:* Data Visualization is a crucial component of data exploration since it aids in accurately understanding the data. Additionally, it aids in your understanding of data's hidden tendencies.
4. *Feature Engineering:* The most crucial step in creating the most accurate predictive model is pinpointing the characteristics (parameters) affecting the forecast or result.
5. *Hyper Parameter:* A parameter whose value is predetermined before the learning process starts is called a hyper parameter.
6. *Cross Validation:* The model evaluation approach known as cross-validation is superior to residuals. The drawback of residual assessments is that they need to anticipate how well a learner will perform when asked to make new predictions for material that it has not previously seen.

3. Simulation Results

Skylearn's accuracy score () function was used to analyze the predictive model's accuracy, and the results showed 72.83% accuracy for about 50,000 invoices. Among all the algorithms used in this paper, XGBoost gives the highest accuracy. In addition, we created the Pearson Correlation of features to provide a better understanding of the model's accuracy for each distinct age bucket. We have also made the distribution plot for the list of features.

The performance metrics such as Mean Square Error, R Square Error and Accuracy are estimated for various machine learning algorithms such as Linear Regression, Linear SVR, Decision Tree Classifier, Random Forest Regressor, XGB Regressor is ahown in Table 2.

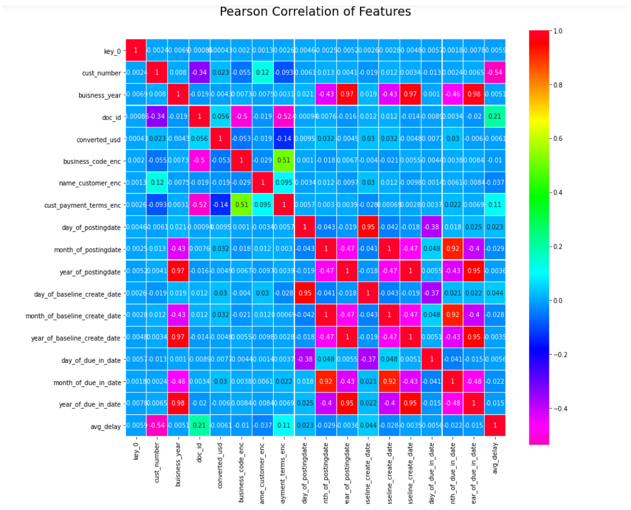


Fig. 1. Pearson correlation using XGBoost

The process of distilling meaningful insights from massive datasets. As we can see above, there is a clear trend in which all of the variables are significantly correlated with one another. Used as a factor in further research. Exploratory component analysis, confirmatory factor analysis, structural equation models, and linear regression all use correlation matrices as inputs and often only account for missing values in pairs. For use as a diagnostic tool with various analyses. In the case of linear regression, for instance, a large number of correlations is indicative of incorrect linear regression results.

Table 2
Comparison of various machine learning algorithms

S.No.	Algorithm	MSE Score	R Square Error	Accuracy
1	Linear Regression	301561975404.23938	0.318425	65.4%
2	Linear SVR	545440475308.176208	-0.232777	66.5%
3	Decision Tree Classifier	158904865010.348816	0.640851	65.9%
4	Random Forest Regressor	155971645907.341858	0.647481	64.4%
5	XGB Regressor	150827747514.634064	0.659107	72.8%

As shown in Figure 1. The Pearson correlation coefficient, a metric for the linear link between two variables, may be used to provide a quantitative assessment of this connection. Its value is between -1 and 1, and it depends on:

- To have a correlation of -1 between two variables shows a totally negative linear relationship.
- When comparing two variables, a value of 0 shows that there is no linear relationship between them.
- An ideal positive linear correlation between two variables would be 1.
- A larger positive value for the correlation coefficient indicates a greater association between the two variables.

Figure 2 represents the distribution plot of "business year Vs density" of the features.

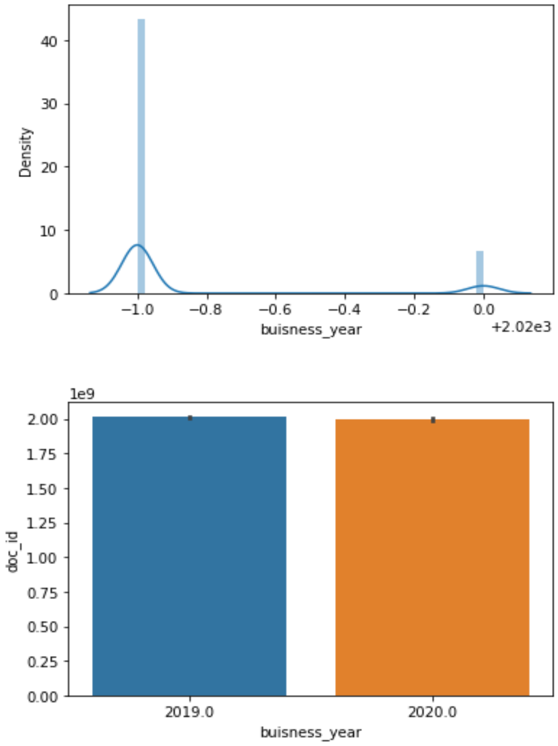
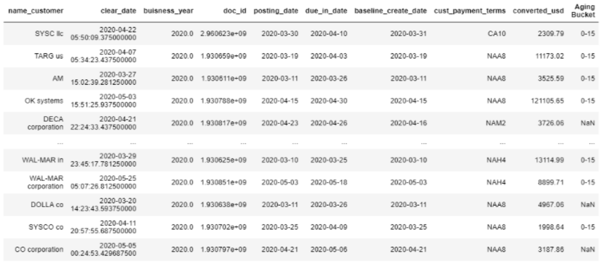


Fig. 2. Comparison of features using XGBoost

Aging Bucket: Aging buckets are intervals during which you can age and examine your debit transactions. For instance, one could create an ageing bucket for all debit items that are one to fifteen days past due as shown in Table 3.

Table 3
Distribution of features



name_customer	clear_date	buisness_year	doc_id	posting_date	due_in_date	baseline_create_date	cust_payment_terms	converted_usd	Aging Bucket
SYSC llc	2020-04-22 05:50:09.375000000	2020.0	2.960623e+09	2020-03-30	2020-04-10	2020-03-31	CA10	2309.79	0-15
TARG us	2020-04-07 05:34:23.437500000	2020.0	1.930659e+09	2020-03-19	2020-04-03	2020-03-19	NAA8	11173.02	0-15
AM	2020-03-27 15:02:39.281250000	2020.0	1.930611e+09	2020-03-11	2020-03-26	2020-03-11	NAA8	3525.59	0-15
OK systems	2020-05-03 15:51:25.937500000	2020.0	1.930788e+09	2020-04-15	2020-04-30	2020-04-15	NAA8	121105.65	0-15
DECA corporation	2020-04-21 22:24:33.437500000	2020.0	1.930817e+09	2020-04-23	2020-04-26	2020-04-16	NAM2	3726.06	NaN
...	...	...	...	...	...	...	...	...	...
WAL-MAR in	2020-03-29 23:45:17.781250000	2020.0	1.930625e+09	2020-03-10	2020-03-25	2020-03-10	NAH4	13114.99	0-15
WAL-MAR corporation	2020-05-25 05:07:26.812500000	2020.0	1.930851e+09	2020-05-03	2020-05-18	2020-05-03	NAH4	8899.71	0-15
DOLLA co	2020-03-20 14:23:43.593750000	2020.0	1.930638e+09	2020-03-11	2020-03-26	2020-03-11	NAA8	4967.06	NaN
SYSCO co	2020-04-11 20:57:55.687500000	2020.0	1.930702e+09	2020-03-25	2020-04-09	2020-03-25	NAA8	1998.64	0-15
CO corporation	2020-05-05 00:24:53.429687500	2020.0	1.930797e+09	2020-04-21	2020-05-06	2020-04-21	NAA8	3187.86	NaN

4. Conclusion

Researchers are encouraged to find financial issues that can be resolved using machine learning approaches, as machine learning is becoming more prominent in the finance industry. The most precious economic asset, accounts receivable, can be fully automated using the way that has been proposed, bringing research one step closer to its goal. The power of business analytics can be naturally augmented by machine learning.

The suggested approach can help to better understand the pattern of bills by anticipating the delay and boosting account receivable collection. Large businesses use this strategy to identify and prioritize past-due invoices to save their collection team's time and money. The proposed XGBoost method provides greater than 72.83% accuracy regarding the payment outcome of an invoice. Therefore, using the suggested strategy in the real world is safe to gain a thorough understanding of emerging trends.

References

[1] V. Mayer-Sch6nberger and K. Cukier, Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt, 2013.

[2] G. Piatetsky-Shapiro, "Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from University to business and analytics," Data Mining and Knowledge Discovery, vol. 15, no. 1, pp. 99-105, 2007.

[3] Shao, M., Zoldi, S., Cameron, G., Martin, R., Drossu, R., Zhang, G., Shoham, D. Enhancing delinquent debt collection using statistical models of debt historical information and account events, US patent 7,191,150, June 2000

[4] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," 2011.

[5] Naoki Abe, Prem Melville, Cezar Pendus, Chandan K Reddy, David L Jensen, Vince P Thomas, James J Bennett, Gary F Anderson, Brent R Cooley, Melissa Kowalczyk, et al. Optimizing debt collections using constrained reinforcement learning. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 75–84. ACM, 2010.

[6] Bart Baesens, Tony Van Gestel, Maria Stepanova, Dirk Van den Poel, and Jan Vanthienen. Neural network survival analysis for personal loan data. Journal of the Operational Research Society, 56(9):1089–1098, 2005.

[7] Butler B. Smith T. Swift T. Williamson J. Scherer W. T. Bailey, D. R. Providian financial corporation: Collections strategy. In Systems Engineering Capstone Conference. University of Virginia, 1999.

[8] Leo Breiman. Random forests. Mach. Learn., 45(1):5– 32, October 2001.

[9] Ricardo Cao, Juan M Vilar, and Andr´es Devia. Modelling consumer credit risk via survival analysis. SORT: statistics and operations research transactions, 33(1):0003–30, 2009.

[10] Michelle LF Cheong and Wen SHI. Customer level predictive modeling for accounts receivable to reduce intervention actions. 2018.

[11] Lore Dirick, Gerda Claeskens, and Bart Baesens. Time to default in credit scoring using survival analysis: a benchmark study. Journal of the Operational Research Society, 68(6):652–665, Jun 2017.

[12] Jerome H. Friedman. Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4):367 – 378, 2002. Nonlinear Methods and Data Mining.

[13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: With Applications in R. Springer Publishing Company, Incorporated, 2014.

[14] Maurice George Kendall. Rank correlation methods. 1948.

[15] Elisa T. Lee and John Wenyu Wang. Statistical Methods for Survival Data Analysis. Wiley Publishing, 4th edition, 2013.

[16] Michal Rychnovsk`y et al. Survival analysis as a tool for better probability of default prediction. Acta Oeconomica Pragensia, 2018(1):34–46, 2018.

[17] Janika Smirnov et al. Modelling late invoice payment times using survival analysis and random forests techniques. Ph.D. thesis, 2016.

[18] Tarun Tater, Sampath Dechu, Senthil Mani, and Chandresh Maurya. Prediction of invoice payment status in account payable business process. In International Conference on Service-Oriented Computing, pages 165–180. Springer, 2018.

[19] Bashar Younes. A Framework for Invoice Management in Construction. Ph.D. thesis, University of Alberta, 2013.

[20] Sai Zeng, Prem Melville, Christian A Lang, Ioana Boier-Martin, and Conrad Murphy. Using predictive analysis to improve invoice-to-cash collection. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1043–1050. ACM, 2008.