REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

REVA Academy for Corporate Excellence (RACE)

Predicting Delays in Invoice Payments using Machine Learning

ARUNA K

SRN: R19MBA54

Date: 27.08.2022

PGDM/MBA in Business Analytics

Capstone Project Presentation
Year: II

race.reva.edu.in

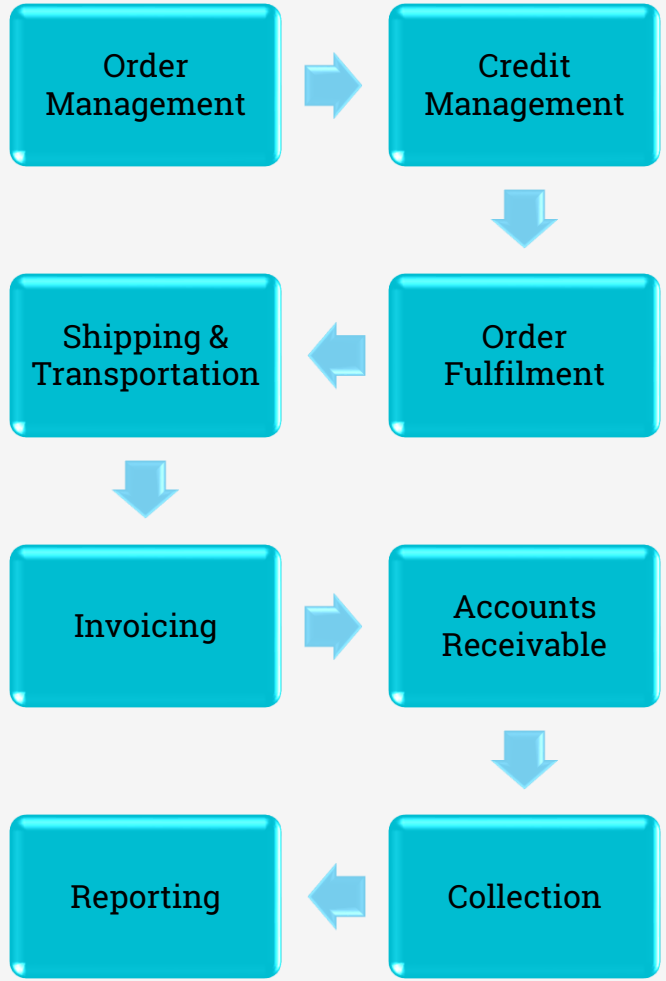
REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Agenda

1. Introduction
2. Accounts Receivables
3. State of Art
4. Problem Statement
5. Project Objectives
6. Project Methodology
7. Business Understanding
8. Descriptive Analytics
9. Modeling
10. Evaluation & Deployment
11. Results & Insights
12. Conclusion & Future Work

Introduction

- Accounts Receivable (AR) are funds the company expects to receive from customers and partners. AR is listed as a current **asset** on the balance sheet.
- Order to Cash (O2C) business process involves AR collections after an invoice is issued to the customer.
- Invoices are used where services and products are provided and they usually contain the rendered charges.
- Typical payment terms provided would be 30, 45, and 60 days to the customer for making full payment of the invoiced amount.
- In business certain customers do not pay on time and intervention actions are required for payment collection.
- To gain an understanding of AR, it is necessary to recognize data patterns in order to forecast whether an invoice will be paid on time or will experience a delay.



O2C Process

Account Receivables and Account Payable

A/R & A/P: Balance sheet categories

Balance Sheet

Assets	Liabilities
What we own	What we owe

+$

Accounts Receivable
Trade Receivables
Debtors
A/R

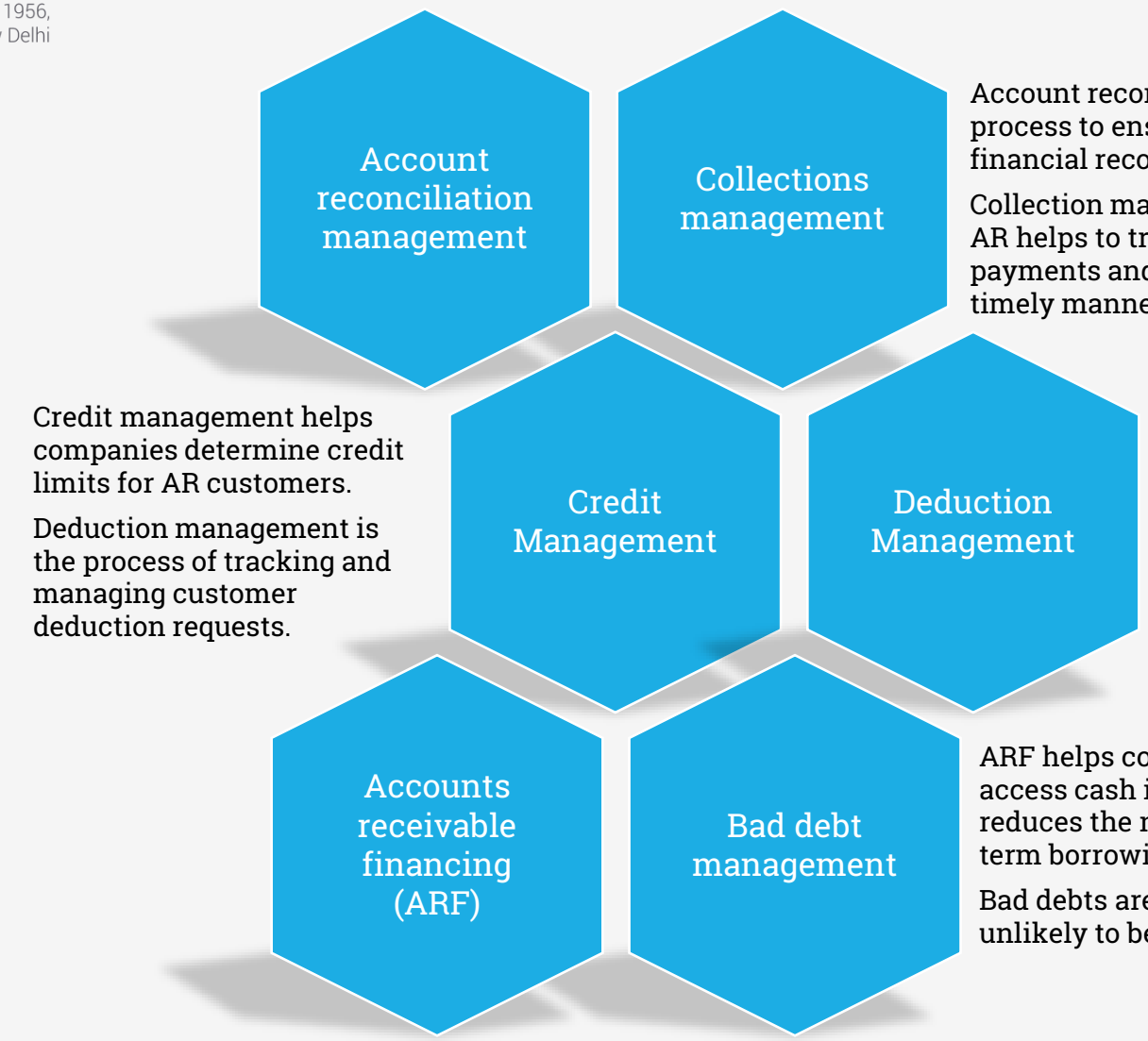
Balance Sheet

Assets	Liabilities
Cash	Payables
Receivables	Accrued Liabilities
Inventory	Debt
Fixed Assets	Equity

-$

Accounts Payable
Trade Payables
Creditors
A/P

AR – Key Processes



Account reconciliation is a process to ensure AR match the financial records.

Collection management process in AR helps to track outstanding payments and collect them in a timely manner.

Credit management helps companies determine credit limits for AR customers.

Deduction management is the process of tracking and managing customer deduction requests.

ARF helps companies to access cash immediately and reduces the need for short-term borrowings.

Bad debts are AR where it is unlikely to be paid.

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Collection Management Process

Identification

Businesses must identify which accounts receivable are overdue and need to be collected.

This step includes reviewing invoices and determining which ones are past due.

Assessment

Businesses must assess the situation of each overdue account and determine the best course of action for collection.

This is a use case where AI / ML can be used to recommend the most appropriate course of action which can result in collection.

Collection

Once a decision has been made on how to collect an overdue account, businesses must take action to actually collect the payment.

This may involve making phone calls, writing letters, or taking legal action.

Disposition

Once an account has been collected, businesses must then dispose of it appropriately.

This may involve filing away records or processing refunds if necessary.

The challenges faced in collection management include timely payment of account receivables, customer relationships, and legal issues.

State of Art

There are few works of literature in the domain of invoice collection prediction. Some interesting papers are going to be introduced now and then used during analysis. To scope the literature, it can be analyzed how different tasks a company would have to tackle to be able to perform prediction of delays.

AUTHORS, PAPER PUBLISHED, YEAR OF PUBLICATION	RESEARCH WORK
Zeng et. al – "Using Predictive Analysis to Improve Invoice-to-Cash Collection", 2018	In this paper the author has gathered invoices from four different firms including 2 fortune 500 companies. several decision tree algorithms were used such as PART and C4.5. The author compares the difference between training a model for each separate firm and training one model on all data and concludes that training the model on combined data from all companies gives a significant improvement in terms of accuracy in all cases. This suggests that invoices sent by different companies (or at least those specific four companies) share similar behavioral patterns.
Hu Peiguang – "Predicting and Improving Invoice-to-Cash Collection Through Machine Learning", 2015	Considering invoice and payment behavior features, the author used multiple classification models for prediction. The Random Forest model had the highest prediction accuracy for predicting if an invoice payment will be on time or delayed, and the delay period. The author concluded that customers with fewer invoices are less likely to have late payments and thus different models have to be built for different customer groups. The author showed that prediction accuracy increases as the number of invoices per customer increases.
Arthur Hovanesyan – "Late payment prediction of invoices through graph features", 2019	The thesis provides a solution to how data about the business supply chains can be used to build a network of SMEs through entity resolution. Furthermore, it shows how this network can be leveraged through methods such as graph embedding, to improve the predictions of late payments. The focus of this thesis is to see whether the addition of features extracted from a graph of related companies can improve the accuracy of late payment predictions.

State of Art (Contd.)

AUTHORS, PAPER PUBLISHED, YEAR OF PUBLICATION	RESEARCH WORK
Tarun Tater et. al – "Prediction of Late Payment of Invoices in Account Payable Business Process", 2018	On the other end of the spectrum, there is a study mentioned in this paper, which focuses on AP. Contrarily to AR, AP is the bills owed by the company to its suppliers for goods and services. The paper discussed the number of paid late invoices that are much smaller in percentage compared to paid on-time invoices in the training data set, hence the classes are imbalanced. The results obtained by training the classifiers show that penalties can be avoided on more than 82% of the invoices being penalized.
Weikun Hu – "Overdue Invoice Forecasting and Data Mining", 2014	The thesis discussed approaches to dealing with imbalanced data, which include 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 demonstrate favorable improvement on classification results.
Yaqi Zhang et. al – "Unstructured Document Recognition on Business Invoice", 2019	In this project, "bags of potential features" are generated to capture aspects of invoice layout, and then evaluated in multiple models to reveal the key properties that identify specific fields of interest.

Considering the learnings from various papers, the study in is focused on classification models using supervised learning with high accuracy in predicting the probability of invoices likely to get delayed thereby providing the organization's business loss due to delayed payments.

Problem Statement

- Every invoice generated must be paid based on the agreement with the customers while booking the orders.
- Invoices that are open will have a 30-day payment term provided to the customer before the collection team intervenes.
- In most cases, payment does not occur prior to the due date, resulting in late payment of the invoice.
- Intervention requires resources and over-intervention could cause unwanted customer dissatisfaction.

Uncertainty whether invoices are paid on time or have delays.

Magnitude of delay – Delayed in number of days.

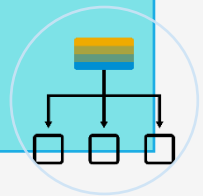
Delays leading to cash traps and decreased cash flow.

Project Objectives

Three major objectives of this study are as follows:

- Whether an invoice will be paid on time or delayed.

Classification



- Prediction of the magnitude of delay in payment showcased in weeks.

Prediction

- Cost savings are beneficial by reducing the number of calls to paying customers with a high likelihood of timely payment.

Cost Savings

Project Methodology

Business Understanding

Invoice processing is a critical part of any business. It's the process of creating, managing, and paying invoices. Businesses can use invoice processing to keep track of customer payments, manage vendor contracts, and streamline their accounting processes.

Data Understanding

The analysis is based on a fortune 500 company's invoice data set that provides technology services. Dataset contains 34,752 invoice data which processes around 1500 invoice transaction data in a month.

Data Preparation

The dataset has been split into training and test data. Dependent variables are "Paid - Ontime" with a Yes/No outcome and "Days Paid Late" indicating the delays in weeks.

CRoss Industry Standard Process for Data Mining Framework

Modeling

Ensemble models such as Random Forest and Extreme Gradient Boosting algorithms have been considered to classify if there is a delay or not and predict the magnitude of delays in weeks.

Evaluation

Evaluation of the model results and review the steps executed to be certain that it properly achieves the business objectives.

Deployment

A web application has been created to show the predicted delays of invoices in weeks for a set of new invoices.

Business Objectives

Business Model Canvas

Key Partners

- ✓ Third Party Solutions provider

Key Activities

- ✓ Invoice Processing
- ✓ Invoice payment collection

Key Resources

- ✓ Collection Teams
- ✓ Finance Team
- ✓ Accounts Team
- ✓ Approver

Value Proposition

- ✓ Timely payment
- ✓ Avoid late fees
- ✓ Improve cash flow
- ✓ Build a vendor-customer relationship
- ✓ Establish trust

Customer Relationships

- ✓ Automated services
- ✓ Ease in payment options

Channels

- ✓ Emails
- ✓ Phone calls

Customer Segments

- ✓ High technology
- ✓ Education
- ✓ Manufacturing
- ✓ Professional Services
- ✓ Government

Cost Structure

- ✓ Infrastructure cost
- ✓ Employees salaries
- ✓ Costs associated with key partnerships

Revenue Streams

- ✓ Invoice Payment

AR Process

Credit Application Process

- ✓ Mechanisms to determine the credit worthiness of the customers
- ✓ Setting credit limits for different types of customers

Send Invoices to Customers

- ✓ Invoices of sales act as a basis for the contract.
- ✓ Invoicing must be done to provide clarity of transactional information to all parties.

Establish Payment terms & Due Dates

- ✓ Most companies choose 30 days as the payment term.
- ✓ Invoices that are pending have to be paid within the established due date.

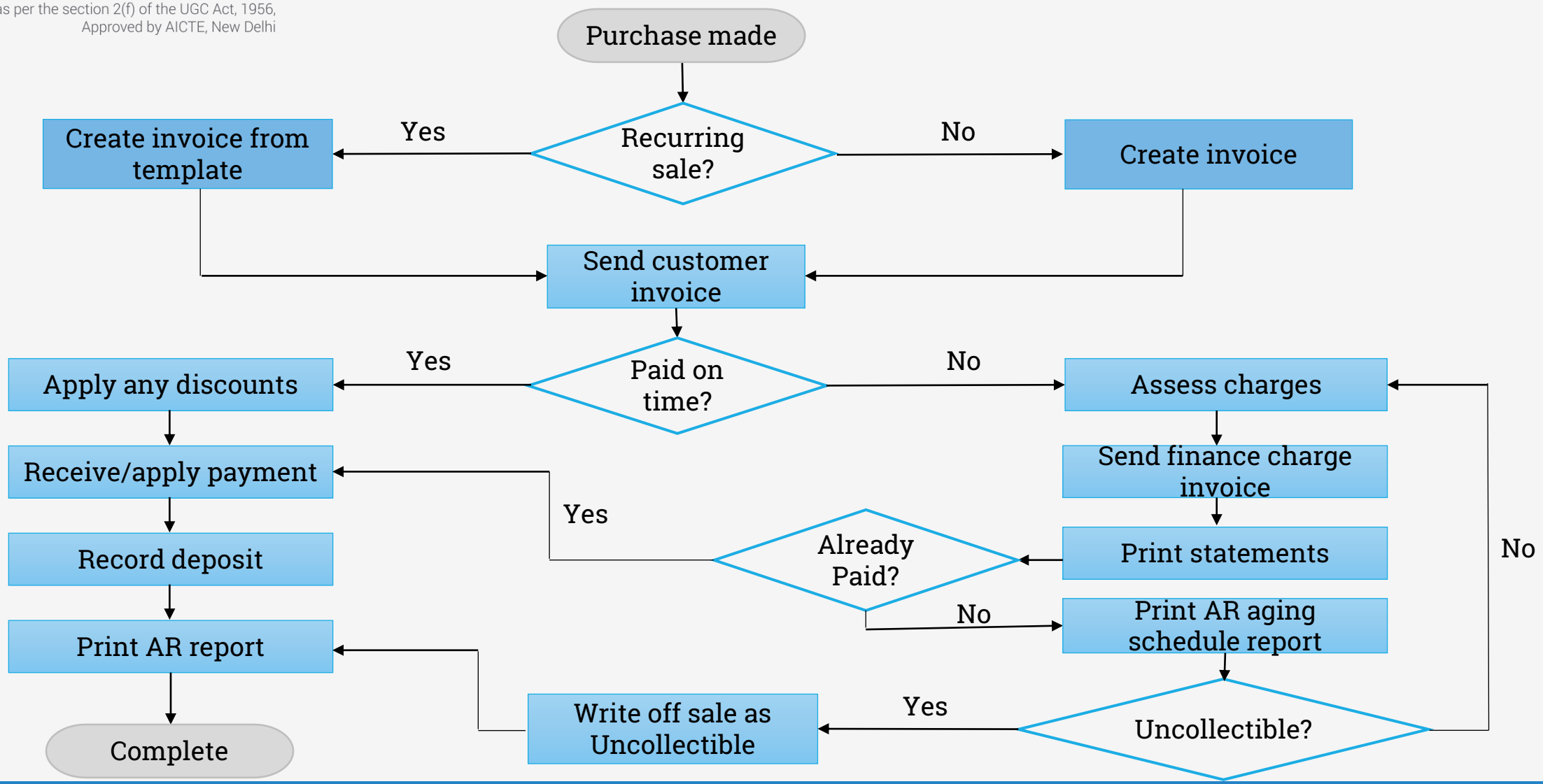
Monitoring & Reporting

- ✓ Outstanding & past due invoices.
- ✓ Recently paid invoices.
- ✓ Balance sheet
- ✓ Reconciliation of assets

Recording AR activity

- ✓ Record all payments received, timestamp of receipt, outstanding debt details, and unpaid invoices.
- ✓ Updates the balance sheet.

AR - Lifecycle



Invoice Processing

Upon receipt of the invoice, the accounting department verifies that the product or service was procured.

Match up the billed amount to the amount that is on the original Purchase Order (PO).

Route the invoice to the right internal employees for approvals.

Resolve any discrepancy in the amounts charged versus the PO.

Set the invoice for payment based on the agreed-upon terms.

Pay the invoice.

Invoice Processing – Challenges

- Impact of late payments on Accounts Receivable
  - ✓ Late payments have a ripple effect on AR which includes reduced cash flow and increased collection costs.
- Increased Day Sales Outstanding (DSO) leads to cash traps
  - ✓ DSO refers to the average number of days a business takes to collect its receivables after a sale.
  - ✓ DSO reduces the return on investment (ROI) and increases borrowing and interest expense to finance.
- Reduced Cash Flow and increased time to collect
  - ✓ Late payments reduce cash flow, negatively impacting the company's ability to pay suppliers, employees, and other operating expenses on time.
  - ✓ Sales orders can sit unprocessed in email inboxes, fax servers, and customer portals.
- Quality of Accounts Receivable
  - ✓ It could lead to an increase in the company's provision for uncollectible accounts and have an unfavorable impact on earnings.
  - ✓ Late payments may ultimately lead to bad debts.

Critical features for Prediction

Data Dictionary – Features considered in the dataset

Features	Description
FY_Quarter	Financial Year with Quarter.
Quarter	Quarter.
Transaction_Number	Combination of unique text_num to identify invoice.
Transaction Type_Mask	Transaction type of the invoice.
Transaction Date	Date on which the invoice is created.
Batch_Source_Masked	Details of Batch Source.
Transaction_Date_Date	Day part from Transaction_Date.
Transaction_Date_Mon	Month part from Transaction_Date.
Transaction_Date_Day	Weekday part from Transaction_Date.
Due_Date	Date by which Invoice must be paid.
Due_Date_Date	Day of the Due Date.
Due_Date_Month	Month of the Due Date.
Due_Date_Day	Weekday of Due Date.
ABE	Accelerated Business Expense.
Credit Hold	Prevent additional credit purchases in case of delay in payment.
PO Mandate	Purchase Order Mandate
Web Invoicing	Scanned invoice sent to customers.
E Invoicing	Invoice transferred between computers
DPLC (days)	Days paid late since due date.
USD_AMT	Cost of the invoice to be paid by the customer.
Paid_15	If the invoice is paid within due date or not.

Key independent variables considered are –

- Paid – Yes/No
- Delays – Number of days

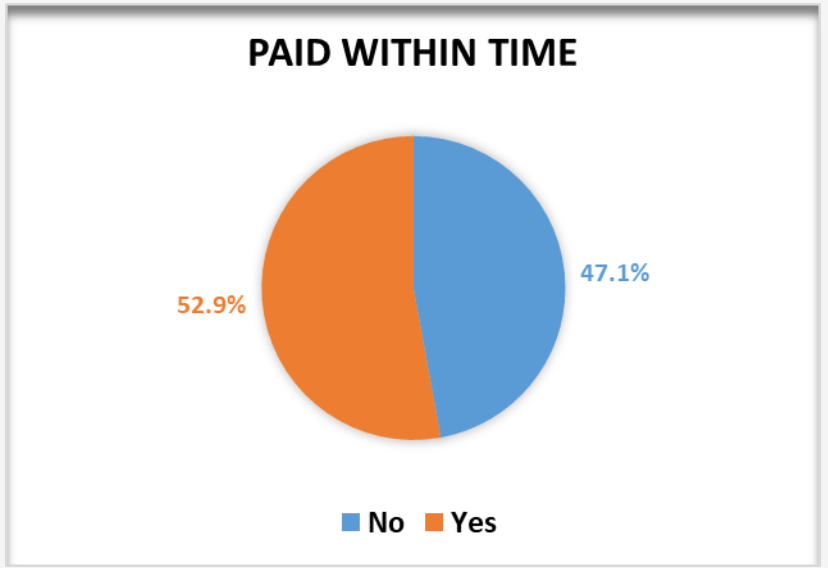
Cost and Time - Delay Analysis

Invoices with delayed payments and the cost associated

Paid on time	Number of Invoices	Cost (Dollars)
No	16,377	$ 84,40,48,460.71
Yes	18,375	$ 72,84,26,849.05
Grand Total	**34,752**	**$ 1,57,24,75,309.76**

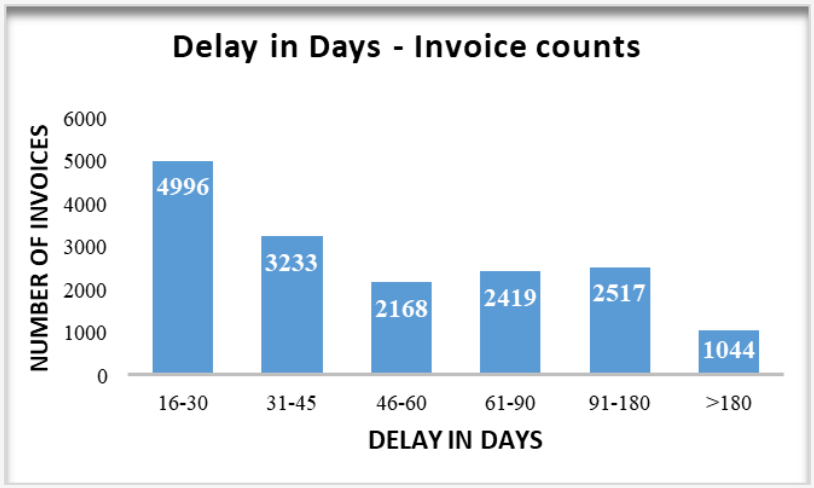
~4 Mn can be saved by the collections team if a per call rate of 0.12 is considered for calling the customers who have delayed payments.

47% of invoices have been delayed for payments.



Invoice Analysis

Delayed invoices bucket (Days)



Over **31%** of invoices are delayed in the delay period of 16-30 days

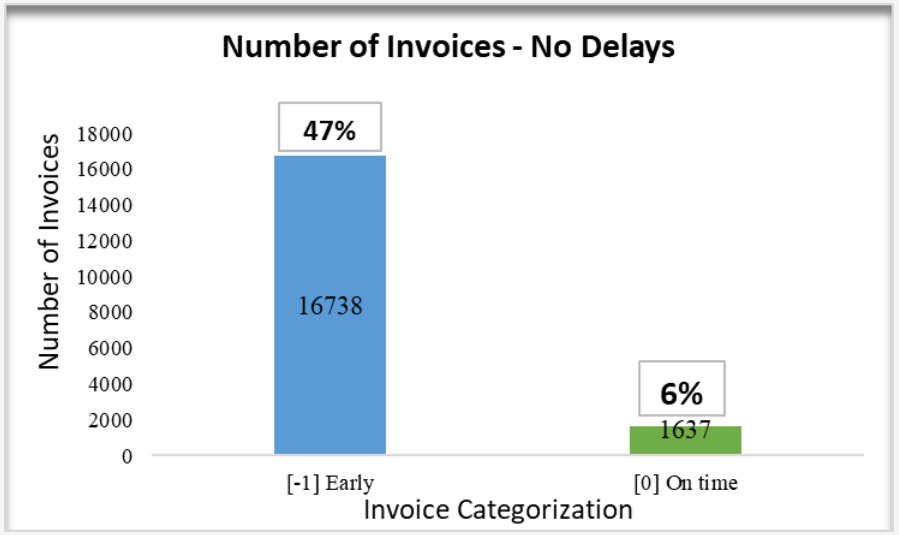
- **64%** of invoice cost can be recovered if the focus is diverted to delays between 16-45 days.

Invoice Cost to be recovered

Delay bucket	Cost (Dollars)	Count of Tickets	%tage of total cost
16-30	33,85,42,215.34	4,996	40%
30-45	19,85,37,481.36	3,233	24%
45-60	11,81,47,058.63	2,168	14%
61-90	9,06,21,887.30	2,419	11%
91-180	7,36,90,188.14	2,517	9%
>180	2,45,09,629.94	1,044	3%
Grand Total	**84,40,48,460.71**	**16,377**	

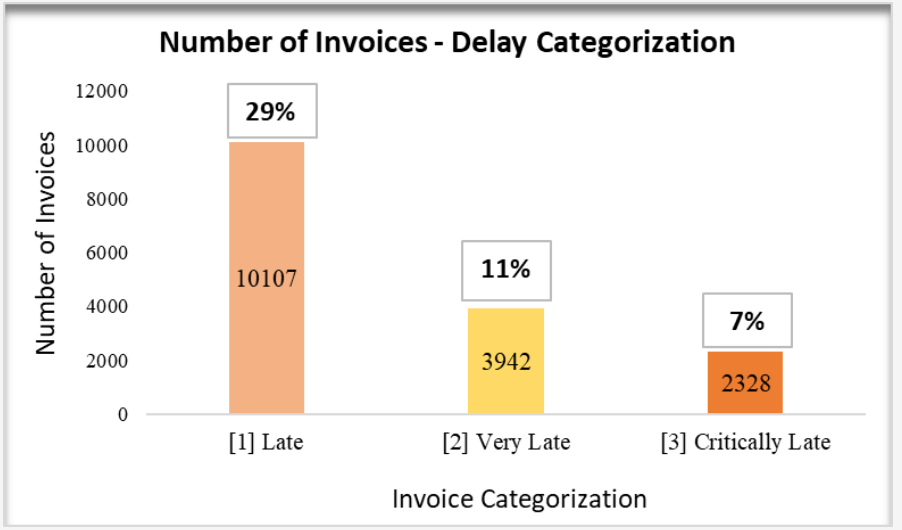
Invoice Categorization

Categorization of invoices – No Delays



Categorizing the invoices if paid early or on time based on the due date.

Categorization of invoices - Delays



Delayed invoices are in turn classified into –

- ✓ late (1) if the payment falls between 1 to 30 days after the due date.
- ✓ Very late (2) if the payment falls late between 31 and 90 days.
- ✓ Critically late (3) if the payment falls late 91 days and onwards.

Data Preparation

Data Type	Time window considered	Number of Invoices (All)
Invoice transaction dataset	January 2018 – September 2020	34,752

Data Preparation Steps:

- Dataset is split into two categories training, and test set.
- Data has been checked for missing values, outlier analysis, and feature engineering.
- The test set is used as the deployment data for the re-validation of the model.
- Independent variables are Paid (classification of invoices with Paid on time = Yes/No) and Delays in days.

Reasons for dropping features

Features	Comments for dropping the features
FY_Quarter	Post factor feature.
Quarter	Post factor feature.
Transaction_Number	Unique Numbers for Customers.
Transaction_Date_Date	Dropped as Transaction Date has been retained.
Transaction_Date_Mon	Dropped as Transaction Date has been retained.
Transaction_Date_Day	Dropped as Transaction Date has been retained.
Due_Date_Date	Dropped as Due Date has been retained.
Due_Date_Month	Dropped as Due Date has been retained.
Due_Date_Day	Dropped as Due Date has been retained.

Modeling

Overview of the dataflow into the Machine Learning Data Model

Following classification algorithms are applied to the invoice dataset:

- Logistic Regression
- XGBoost Classification
- XGBoost Regression
- Random Forest Classification
- Random Forest Regression
- Neural Network
- K-Neighbors
- Linear SVC

Predicted Features –

Paid – Delay (Yes/No): Classify in case of invoice whether the cost will be paid on time or delayed.

Days paid late: Predict the magnitude of delay in payment in weeks.

Model Evaluation

- Multiple classification models have been considered to compare the performance. The accuracy metric has been indicated in the table for all the models.

- Ensemble models such as Extreme Gradient Boosting (XGBoost) and Random Forest have been used in the study.

- The AUC curve for XGBoost Classification was 0.74 and for Random Forest Classification it came out to 0.73.

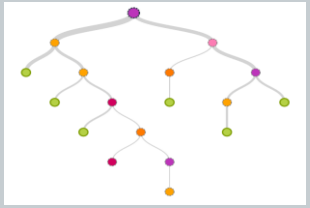
Models	Accuracy
Logistic Regression	68.62%
XGBoost Classification	**81.57%**
Random Forest Classification	**80.60%**
Neural Network	36.73%
K-Neighbors	45.02%
Linear SVC	55.96%

- XGBoost Classification and Random Forest Classification models have a higher accuracy score close to 82% and 81% respectively and hence these models have been considered for deployment.

- Hyper parameter tuning was performed on the XGBoost model by toggling the values from 100 to 1000 for ‘n_estimators’ which represents the number of trees. There was a slight improvement in the accuracy.

Model Deployment

Data Preparation & ML Model



Web Application Development

python

Flask

Deployment

Pythonanywhere

pythonanywhere

End Users

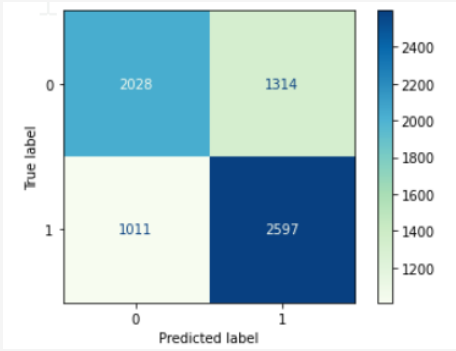
Results and Insights

- The models are evaluated using the confusion matrix and the model performance is calculated using the True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) values.

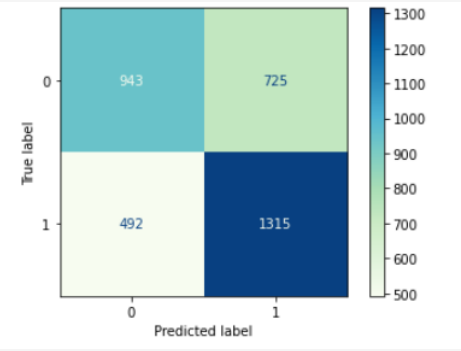
- Confusion Matrix indicates for XGB indicates –
  - ✓ 2597 (TP) invoices have been predicted as delays which are actual delays.
  - ✓ 2028 (TN) invoices have been classified as no delays and are not delayed in reality.
  - ✓ 1314 (FP) invoices have been classified as delays but are not delayed.
  - ✓ 1011 invoices (FN) have been classified as no delays but they are delayed in actuality.

Metrics/Models	XGBoost Classifier	Random Forest Classifier
Precision	69.40%	68.63%
Recall	77.91%	78.34%
F1 Score	69.78%	69.23%
AUC	74.69%	73.67%

Metrics for Invoice Classification



Confusion Matrix - XGB



Confusion Matrix - RF

Conclusion

- Successfully used predictive models to classify if a customer will pay the outstanding invoices on time with a high probability and predicted the magnitude of delays.
- The outcome of this study is an attempt to reduce intervention actions taken on the customers, thus reducing costs and improving customer relationships.
- The project focused on the invoice level details by including features such as due date, cost of the invoice along with various other features.

Future Work

- This project does not capture customer-level data. In the future study, customer related data such as customer billing cycle, region wise data, and employee level data shall be included.
- Since the scope of the data was at invoice level, not many variables were available for seasonality pattern study. This will also be considered in phase 2 of this project.
- There are some invoices that are overdue for more than a few months which are technically in dispute. So, identify the invoice which is likely to get into any sort of dispute.

GitHub Link

GitHub Link Contains:

- Code
- HTML Files for Pandas Profiling
- HTML File for Code Output

GitHub Link :
https://github.com/ArunaAI23/Prediction-of-Delays-in-Invoice-Payments-using-Machine-Learning.git

References

1. Appel, A. P., Malfatti, G. L., Cunha, R. L. de F., Lima, B., & de Paula, R. (2020a). *Predicting Account Receivables with Machine Learning*. http://arxiv.org/abs/2008.07363

2. Bachelor, W. H., Simchi-Levi, D., Donald, H. N., Harleman, M., & al Engineering, E. (2016a). *Overdue Invoice Forecasting and Data Mining Signature redacted ... red.acted Chair, Graduate Pro am Committee*.

3. Breiman, L. (2001). *Random Forests* (Vol. 45).

4. Brownlee, J. (2016). *A Gentle Introduction to XGBoost for Applied Machine Learning*.

5. Ezvan, J.-L., & Girard, F. (2018). *University Paris-Dauphine Predicting late payment of an invoice*.

6. Fernandez, E. B., & Yuan, X. (n.d.). *An Analysis Pattern for Invoice Processing*.

7. Hovanesyan, A. (2019). *Late-payment prediction of invoices through graph features*.

8. Investopedia. (2021).

9. Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

10. Korotina, A., Mueller, O., & Debortoli, S. (2015). *Association for Information Systems AIS Electronic Library (AISeL) Real-time Business Process Intelligence. Comparison of different architectural approaches using the example of the order-to-cash process*. http://aisel.aisnet.org/wi2015/114

11. Kuhn, M., & Johnson, K. (2013). Over-Fitting and Model Tuning. In M. Kuhn & K. Johnson (Eds.), *Applied Predictive Modeling* (pp. 61–92). Springer New York. https://doi.org/10.1007/978-1-4614-6849-3_4

12. Li, Ying., ACM Digital Library., Association for Computing Machinery. Special Interest Group on Knowledge Discovery & Data Mining., & Association for Computing Machinery.

13. Special Interest Group on Management of Data. (2008). *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM.

14. Lopes, S., & Rebelo, C. (2021). *Predicting Account Receivables Outcomes with Machine-Learning*.

15. Peiguang, H. (2015). *Predicting and Improving Invoice-to-Cash Collection Through Machine Learning*.

References (Contd.)

16. Pfohl, H. C., & Gomm, M. (2009). *Supply Chain Finance: Optimizing Financial Flows in Supply Chains*.

17. Quiry, P., Vernimmen, P., Fur, Y., Dallocchio, M., & Salvi, A. (2009). *Corporate Finance: Theory and Practice*.

18. Ramanei, T. a. -p., Abdullah, N. L., & Khim, P. T. (2021). Predicting Accounts Receivable with Machine Learning: A Case in Malaysia. *2021 International Conference on Information Technology (ICIT)*, 156–161. https://doi.org/10.1109/ICIT52682.2021.9491773

19. Rebelo, S. L. da C. (2022). *Predicting Account Receivables Outcomes with Machine-Learning*.

20. Shah, H. S. (2016). Licensed Under Creative Commons Attribution CC BY Customer Payment Prediction in Account Receivable. *International Journal of Science and Research (IJSR) Index Copernicus Value*, 7–296. https://doi.org/10.21275/ART20194177

21. Smirnov, J. (n.d.). *Modelling Late Invoice Payment Times Using Survival Analysis and Random Forests Techniques*.

22. Stahlbock, R., Weiss, G. M., & Abou-Nasr, M. (2018). *Proceedings of the 2018 International Conference on Data Science : ICDATA '18*.

23. Tarawneh, A. S., Hassanat, A. B., Chetverikov, D., Lendak, I., & Verma, C. (2019). Invoice Classification Using Deep Features and Machine Learning Techniques. *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, 855–859. https://doi.org/10.1109/JEEIT.2019.8717504

24. Tater, T., Dechu, S., Mani, S., & Maurya, C. (2018). Prediction of invoice payment status in account payable business process. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11236 LNCS*, 165–180. https://doi.org/10.1007/978-3-030-03596-9_11

25. Zeng, S., Melville, P., Lang, C. A., Boier-Martin, I., & Murphy, C. (2008). Using Predictive Analysis to Improve Invoice-to-Cash Collection. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1043–1050. https://doi.org/10.1145/1401890.1402014.

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Annexure

Plagiarism score

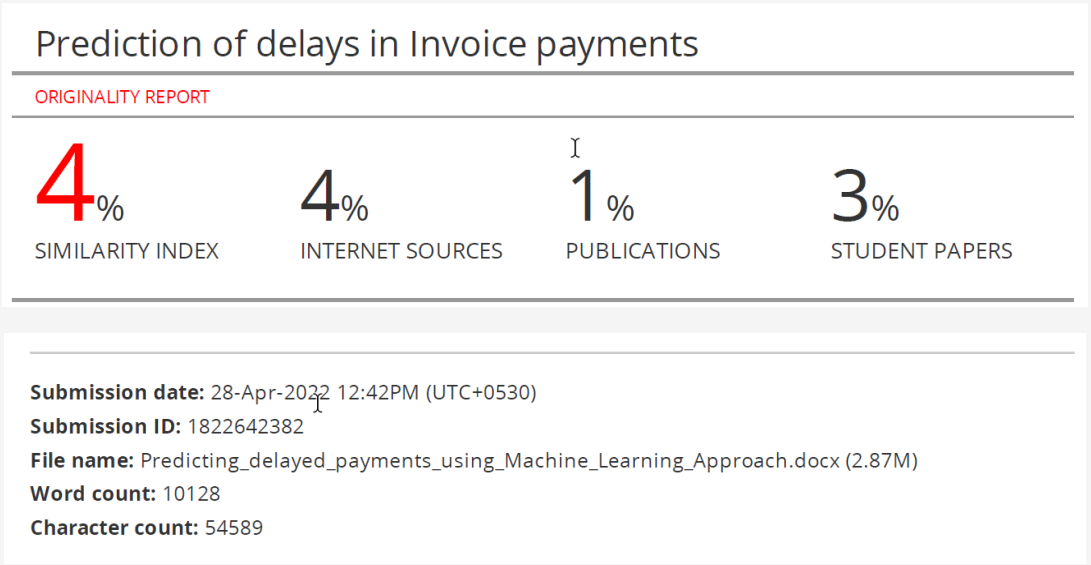
This is to certify that this project report titled **Predicting Invoice Payments Delay using Machine Learning Approach** was scanned for similarity detection. Process and outcome are given below.

Software Used: **Turnitin**

Date of Report Generation: **12.08.2022**

Similarity Index in %: **4%**

Total word count: **10128**


Prediction of delays in Invoice payments

ORIGINALITY REPORT

4%	4%	1%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

Submission date: 28-Apr-2022 12:42PM (UTC+0530)
Submission ID: 1822642382
File name: Predicting_delayed_payments_using_Machine_Learning_Approach.docx (2.87M)
Word count: 10128
Character count: 54589

REVA
UNIVERSITY

Bengaluru, India

Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Thank You!