REVA UNIVERSITY
Bengaluru, India

A Project Report on

Prediction of Delays in Invoice Payments using Machine Learning

Submitted in Partial Fulfilment for Award of Degree of

Master of Business Administration in Business Analytics

Submitted By

Aruna Kashinath

R19MBA54

Under the Guidance of

Mithun D J

Senior Manager – Data Science

RACE, REVA University

REVA Academy for Corporate Excellence - RACE

REVA University

Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064

race.reva.edu.in

August, 2022

Candidate's Declaration

I, **Aruna Kashinath** hereby declare that I have completed the project work towards the Second year of Master of Business Administration in Business Analytics at, REVA University on the topic entitled **Prediction of Delays in Invoice Payments using Machine Learning** under the supervision of **Mr. Mithun D.J.** This report embodies the original work done by me in partial fulfilment of the requirements for the award of a degree for the academic year **2022.**

Place: Bengaluru

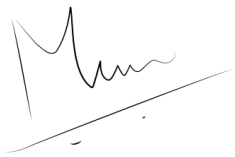
Date: 27th August 2022

Name of the Student: **Aruna Kashinath**

Signature of Student

Certificate

This is to Certify that the project work entitled **Prediction of Delays in Invoice Payments using Machine Learning** carried out by **Aruna Kashinath** with SRN R19MBA54**,** is a bonafide student at REVA University, is submitting the Second-year project report in fulfilment for the award of **Master of Business Administration** in Business Analytics during the academic year 2022. The Project report has been tested for plagiarism and has passed the plagiarism test with a similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the said degree.



Signature of the Guide

Name of the Guide **Mr. Mithun D.J.**

Signature of the Director

Name of the Director

Dr. Shinu Abhi

External Viva

Names of the Examiners

1. Vaibhav Sahu, Strategic Cloud Engineer, Google
2. Abhishek Sinha, Data Science Manager, Capgemini

Place: Bengaluru

Date: 27th August 2022

Acknowledgment

I am highly indebted to **Dr. Shinu Abhi**, Director, and Corporate Training for their guidance and constant supervision as well as for providing necessary information regarding the project and also for their support in completing the project.

I would like to thank my project guide **Mr. Mithun D. J.** for the valuable guidance provided to understand the concept and execute this project. It is my gratitude towards **Dr. Jay Bharateesh Simha** and all other mentors for the valuable guidance and suggestion in learning various data science aspects and for their support. I am thankful to my classmates for their aspiring guidance, invaluable constructive criticism, and friendly advice during the project work.

I would like to acknowledge the support provided by Hon'ble Chancellor, **Dr. P Shyama Raju**, Vice Chancellor, **Dr. M. Dhananjaya,** and Registrar, **Dr. N Ramesh**. It is sincere thanks to all members of the program office of RACE who were supportive of all requirements from the program office.

It is my sincere gratitude towards my parents, and my family for their kind co-operation and encouragement which helped me in the completion of this project.

Place: Bengaluru
Date: 27th August 2022

REVA UNIVERSITY
Bengaluru, India

Similarity Index Report

This is to certify that this project report titled **Prediction of Delays in Invoice Payments using Machine Learning** was scanned for similarity detection. Process and outcome are given below.

Software Used: **Turnitin**

Date of Report Generation: **28-Apr-2022**

Similarity Index in %: **4%**

Total word count: **10128**

Name of the Guide: **Mr. Mithun D J**

Place: Bengaluru

Date: 28.04.2022

Name of the Student: Aruna Kashinath

Signature of Student

Verified by: Dincy Dechamma

Signature

Dr. Shinu Abhi,

Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	CRISP-DM	Cross-Industry Process for Data Mining
2	EDA	Exploratory Data Analysis
3	AR	Accounts Receivable
4	LR	Logistic Regression
5	SVC	Support Vector Classification
6	RF	Random Forest

List of Figures

No.	Name	Page No.
Figure 1.1	Procure to Pay Cycle	13
Figure 5.1	"CRISP-DM Framework"	23
Figure 6.1	O2C and I2C processes	25
Figure 6.2	Algorithmic Solution for Delayed Invoice Payments	26
Figure 7.1	Sample Data in masked format	29
Figure 7.2	Describing the data - ABE	30
Figure 7.3	Describing the data – PO Mandate	30
Figure 7.4	Describing the data – Credit Hold	30
Figure 7.5	Describing the data – Web Invoicing	31
Figure 7.6	Describing the data – E-Invoicing	31
Figure 7.7	Classification of invoices in terms of delay	32
Figure 7.8	Delay buckets in Days	33
Figure 7.9	Invoice Delay vs No Delay	34
Figure 7.10	Invoice Categorization – No Delay Payments	35
Figure 7.11	Invoice Categorization – Delay Payments	35
Figure 7.12	Delayed Invoice Summary	36
Figure 7.13	Number of Invoices created in 2018	37
Figure 7.14	Number of Invoices created in 2019	37
Figure 7.15	Number of Invoices created in 2020	37
Figure 7.16	Sample Payment Term	38
Figure 7.17	Invoice amount and delay or not	39
Figure 7.18	Invoice amount and delay level	39
Figure 7.19	Average amount of delayed invoice versus delayed days	40
Figure 8.1	Data Extraction Flow	41

Figure 10.1	Confusion Matrix of Invoice Delay Prediction for test result	52
Figure 10.2	ROC Curve	53
Figure 12.1	Delay in weeks buckets for XGBoost Regressor Output invoices	57

List of Tables

No.	Name	Page No.
Table 6.1	Cash Inflows and Outflows examples	24
Table 7.1	Data Dictionary	28
Table 8.1	Reason for dropping features	41
Table 8.2	Key attributes of the data	44
Table 8.3	List of New Features	46
Table 9.1	Confusion matrix metrics for binary classification of invoices	47
Table 10.1	Metrics for Paid label	52
Table 10.2	Metrics for DPLC (Delay in weeks) label	52
Table 12.1	XGBoost Classifier Prediction	56
Table 12.2	Random Forest Classifier Prediction	56

Abstract

Accounts Receivable (AR) is the profitable asset of an organization and can transcend to financial difficulties for firms if not managed efficiently. In order to gain an important understanding of AR, data patterns must be recognized to forecast the likelihood of an invoice being paid on time or having delays in payment.

An invoice is created for a customer with the amount that is owed to the supplier after receiving goods or services. Yet this may not usually occur prior to the due time of an invoice transaction, implying an invoice transaction cost is frequently not paid on time.

This project aims at predicting the delay in payments well in advance, for customer invoices that would help the Collections Management team in proactively identifying such accounts and taking necessary actions against open or unpaid invoices. The data is mostly in a structured format capturing the invoice payment details.

Historical data on previous invoices from Q1 of FY2018 to Q3 of FY2020 form the basis of our study on factors that are most significant in late payments.

The project exemplifies a machine learning forecasting prediction that can be used to construct a model for forecasting the cost-benefit for an invoice transaction due (open) as per the past data. The suggested technique forecasts the delays in terms of accuracy and is implemented as per the backdrop of AR data in real life. At last, outcomes have been showcased in terms of delay in weeks which proves that categorization of delays in invoice payments can save a significant amount of collection time.

The Business Impact of this project would result in the Collection Management team reaching out to respective customers or account holders within the period when the invoices are open, thereby reducing the accounts receivables for the organization.

Keywords: Invoice Processing, Invoice Payment, Delayed Invoices, Accounts Receivable, Predictive Modeling

Contents

List of Abbreviations 6
List of Figures 6
List of Tables 7
Abstract 8
Chapter 1: Introduction 12
Chapter 2: Literature Review 16
Chapter 3: Problem Statement 20
Chapter 4: Objectives of the Study 21
Chapter 5: Project Methodology 23
Chapter 6: Business Understanding 24
Chapter 7: Data Understanding 28
7.1 Collecting the Initial data 28
7.2 Describing the data 29
7.3 Target Setting 31
7.4 Exploratory Data Analysis (EDA) 33
Chapter 8: Data Preparation 41
8.1 Selecting data and dropping duplicate rows 42
8.2 Treating Missing Values 42
8.3 Checking for Outliers 43
8.4 Feature Extraction: 43
8.4.1 Choosing the right data 44
8.4.2 Two levels of features 44
8.4.3 Extra information and unexpected features 45
8.4.4 New features extracted for learning 46
Chapter 9: Modeling 47
9.1 Evaluation metrics 47
9.2 Machine Learning Algorithms for Supervised learning 48
Chapter 10: Model Evaluation 52
Chapter 11: Deployment 55

Chapter 12: Analysis and Results 56
Chapter 13: Conclusions and Recommendations for future work 58
Bibliography 59
Appendix 62
Plagiarism Report 62
Annexure 65

Chapter 1: Introduction

Every Business aims at ways to contain costs and increase the cash inflow. To achieve this the financial system of the organization must work with great efficiency. Within this financial system, the ***Accounts receivable*** classify the costs to be paid by the clients or account holders.

Every invoice that is generated, must be paid based on the agreement with the customers while booking the orders. Open invoices will have a 30-day payment term provided for the customers to complete the payments and 45 days limit has been sanctioned prior the interference from the vendor or supplier organization. It is a common problem among many organizations that customers fail to pay on time, and this results in subsequent follow-ups to remind their customers to pay the outstanding invoice amount (Stahlbock et al., 2018).

When the number of invoices generated per day is huge, manually identifying the late payments or debts becomes critical for the AR team. To manage this process and to reduce the accounts receivables for the organization, the collection management team reaches out to its customers regarding their due payments to the organization.

Prior to Machine Learning is in place, organizations created visualizations that helped to take business-related decisions manually, called descriptive analysis. The project here demonstrates the implementation of a forecasting study on Account Receivable which is the most valued asset of a business. (Shah, 2016)

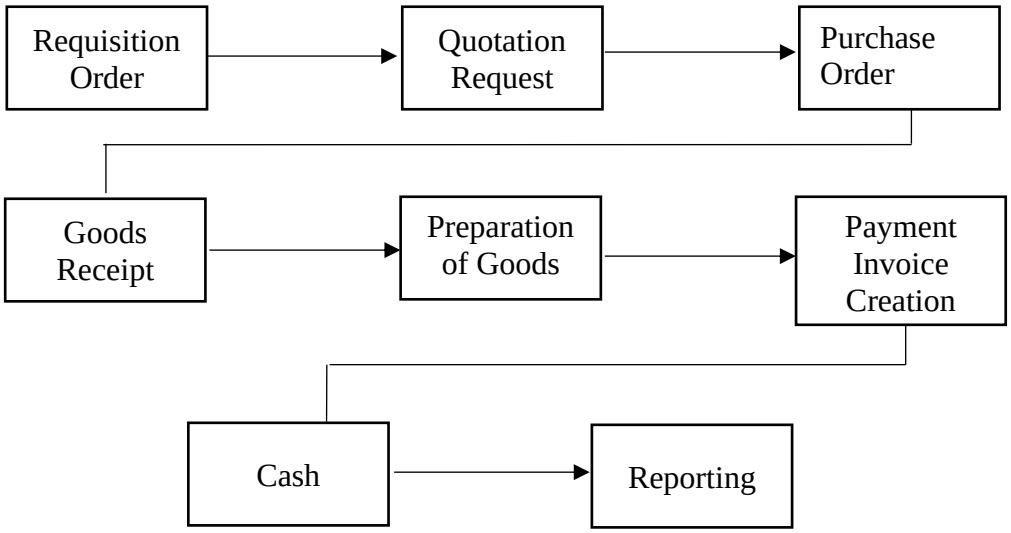


Figure 1.1 Purchase to Payment Cycle

A seamless and effective Procure to Pay (P2P) process defined in Figure 1.1 is core to the growth of an organization. Having a smooth and organized procure-to-pay process will optimize business expenditures and ensure that inventory orders are processed promptly.

The procurement and payment cycles are comprised of several steps involving procurement and financial procedures.

Requisitioning

- A business user communicates the need for some goods or services to the procurement team.

Review and Approval

- Specifically, the approving manager reviews the requirement and evaluates its cost implications and purchase justifications.

Request for Quotation (RFQ)

- Organization requests a quote from a supplier to purchase specific products or services.

Creation of Purchase Order (PO)

- The purchasing department has received a formal purchase order for valid purchase request.

Good Receipt

- Supplier provides the requested goods or services, and the necessary goods or service receipts are created.

Creation of payment invoice and approval

- The invoice should be approved if there are no discrepancies between the PO and the goods receipt, the invoice should be approved and forwarded to accounts payable.

Payment

- The finance team can release funds and pay the vendor according to the terms of an approved invoice.

Reporting

- It will provide a better understanding of how metrics can be used to emphasize the positive and negative aspects of this process and identify any opportunities for improvement in the future.

Benefits of P2P cycle are optimizing procurement processes, reducing processing costs, improve supplier relationships, simplified data reporting and leveraging negotiations.

For businesses, AR are usually collected once the product or a service invoice is sent to a consumer or client. Generally, the terms of payment for the customer are 30, 45 or 60 days for total cost of the invoice transaction. There are several industries that faces same issue of not receiving payments from their customers and having to take action to remind them to pay their outstanding invoices. Interventions of this type involve an expenditure of time and resources, and they are likely to result in poor customer satisfaction.

There are several unintended consequences of weak accounts receivable management processes. Despite the fact that most companies use ERP/accounting systems to manage their AR, this process still requires a significant amount of manual effort and time to manage. As a result of these manual processes, the collection team must spend considerable time finding customer data, resolving data errors, updating spreadsheets, and performing other tasks. They should spend more time engaging with customers, resolving disputes, and undertaking other activities that can speed up payments.

In the project demonstrated below, the focal point is bettering the efficacy of AR. Data from the past year's transactions have been utilized to identify patterns and create a predictive model. Using this model, new invoices will be categorized into payment buckets. Having this information can allow the collection department to concentrate on offending invoices along with bringing in proactive steps for speeding the invoice payment collection process.

Chapter 2: Literature Review

For this study, about 20 research papers have been reviewed under various topics related to invoice processing and the different methodologies used to analyse the estimation of consumer payment in AR. Some papers have been discussed herein.

The Order-to-Cash (O2C) explains a combined method of business encompassing major steps involved in completing an order for a service, from order entry to receipt of payment. Although the nature of the processes depends on the nature and magnitude of the organization, most O2C processes are built on a much alike workflow. As a central part of the collections process, it deals with concentration on account, contact point of a consumer, collection calls, and escalation. Consumers are generally reached at planned time period, however, in case if a consumer is reached out late then the chances of the invoices being paid on time are reduced. Repeated follow-ups with "good" customers can yield a reduced customer satisfaction (Li et al., 2008).

Managing receivables is an integral part of the credit-to-cash transformation sequence, which includes collections, management of payments, and the aging of debtors. Management of AR may not be transparent, resulting in long aging borrowed due to ineffective management of the collection (Ramanei et al., 2021).

Recovering debts contend a predominant part of the business's finances. If invoice payments are delayed, this is the origin of issues pertaining to the company's liquidity and cash flow forecast. The financial stability of businesses is the outcome of accurate forecasting of cash flows. An experimental analysis of the payment behavior of debtors is conducted in this thesis and results for forecasting analysis are found to provide useful information (Smirnov, n.d., 2016).

Account receivables are one of the main challenges in business operations. When it comes to invoices and cash collection, the process has been inefficient,

and a growing number of unpaid invoices leads to issues with the flow of cash due to the accumulation of overdue invoices. Additionally, there is a discussion of how to deal with data imbalance, which also includes techniques for sampling and measurements of performance.(Bachelor et al., 2016a).

"It is commonly agreed that account receivable is the most valuable asset of any business firm. It can be a source of financial difficulties for firms when they are not efficiently managed and underperforming. So, it is important to identify data patterns in account receivable and get meaningful insight from account receivable data" (Shah, 2016).

There are many pragmatic and notional studies available in the literature on improvising the Accounts Receivables process. There are several phases to this pattern, including the creation, validation, and payment of an invoice. Invoice creation and payment are described by two simple patterns (Fernandez & Yuan, 2009).

The challenge of invoice identification lies in its ability to extract structured information from unstructured documents with unforeseeable formats and phrase. One of the studies by Yaqi Zhang, Billy Wan, and Wenshun Liu suggests various prospective attributes are brought out to capture particulars of invoices, estimated in multiple models to disclose key characteristics (Zhang et al., 2016).

A strenuous task is determining whether invoices are likely to be delayed because of huge capacity of invoices along with the classification. This thesis focused a binary classification using supervised learning by considering the details of the invoice transaction along with actions taken on the invoice to forecast in case of a delay and flag it for balance (Tater et al., 2018).

In this research, the aim is to provide a prediction model tailored for every client. Different models were trained to use a prediction algorithm that trains quickly while still being accurate (Ezvan & Girard, 2018).

This research uses machine learning to forecast good accuracy in invoice statuses, there has been consideration of past data and transient attributes to refine the accuracy of models, along with comparing outcomes. In order for collectors to rank consumers, it is mandatory to consider not only the payment cost overdue, but also the likelihood of delays (Appel et al., 2020a).

"We are interested in improving AR collection through machine learning for three reasons. First, AR collection can easily be a source of financial difficulty for firms, if not well managed. It is, therefore, of great interest to manage it more effectively. Also, most of the AR collection actions nowadays are still manual, generic, and expensive. For instance, it seldom takes into account customer specifies, and neither has any prioritizing strategies.
Last and most importantly, commercial firms now are accumulating a large amount of data about their customers, which makes the large-scale data-driven AR collection possible" (Peiguang, 2015).

In this study, they have proposed a mechanized viewpoint to segregate invoices into three types: handwritten, machine-printed, and receipts. They have put forward a method based on AlexNet which is a deep convolutional neural network that can be used to extract features (Tarawneh et al., 2019).

In the thesis submitted by Arthur Hovanesyan, in order to improve the prediction of late payments, there exists a solution that uses data about business supply chains to build a network of SMEs through entity resolution and shows how this network can be applied through embedding graphs methodology. Also, the thesis focuses on the inclusion of the attribute that has been brought out from a graph of interconnected organizations with the possibility of improving the accuracy of forecasting delayed payments (Hovanesyan, 2019).

In the paper, a fresh tribune for tax-control network invoice machine management has been divulged. By adapting computer and network technology,

it permits tax authorities to accumulate and manage taxation details further skillfully (Zhu et al., 2016).

Chapter 3: Problem Statement

The Collection Management team is responsible for payment collections against open invoices. When an order is made and an invoice is generated, the customer has a 30-day time period to make the payment. The company also provides a grace period of 15 days from the due date. Failure of payment within this period is considered as late payment.

There can be valid delays from the organization's end such as:

- Product not being delivered on time OR some customers order multiple products that may reach in batches.
- Specific factors such as time (year closure), LOB, order type have delayed payment.
- Some customers like to have e-invoices instead of the physical invoice which may lead to delays.
- Purchase Order (PO) mandate being yes indicates the customer has mandate PO.

Predicting the delay in invoices – delay along with the delay period (in weeks) using appropriate Machine Learning Algorithms and to explore the key drivers leading to delay in payments. The purpose is to help the collections team proactively reach out to these customers, resolve the issues, and complete the payments.

Chapter 4: Objectives of the Study

The scope of this study is to have a concrete way of proactive prediction/classification of invoices that can end up having delayed payments. The advantage is that the collections team can plan their follow-up process well in advance for those invoices that can have delay in payments.

Companies use Order-to-Cash (O2C) to process and receive orders from the customer. If not well managed, Collection of AR is the origin of financial hardship for companies. Effective process is important to the success of finance management of a company. Furthermore, AR collection is done manually, in a collective fashion, and at a high cost. There are no prioritizing strategies, nor it considers customer specifics. The fact that companies now collect a lot of data regarding their clients can help driving intensive, an information-based approach ensuring AR collections achievable.

Using an AI/ML approach to identify well in advance which invoice is most likely to get delayed and provide solutions to operations. To determine factors that are more significant in terms of identifying the payment delay.

Some factors are evident, such as - Transaction type, LOB, a certain period of years, geography of customers, the value of the invoice, and due date. There might be some hidden factors that need to be identified.

Building supervised learning models to forecast overdue invoices in advance can help a 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.

Machine-learning models to forecast the payment process of invoices have been developed recently and few studies have been published on this topic. Supervised learning has been implemented in the improvement of the invoice to the cash collection process. The models demonstrate high prediction accuracy

for delayed invoice payment, which provides an excellent framework for machine learning studies on invoice collection.

There are three reasons why it is crucial to improve AR collection through machine learning:

- The collection of AR may easily cause financial difficulties for firms if it is not managed properly.

- Majority of AR collection actions are still performed manually, generically, and cost prohibitive. It seldom considers customer specifics and has no prioritization strategy.

- A growing number of commercial enterprises collect a large amount of data about their customers, making it possible to collect large-scale AR data using data-driven methods.

Machine learning can be used to accomplish the following objectives:

1. *Classification of invoices – which invoices will have a delay in payment*
2. *Probability/prediction in weeks by which the invoice will be paid to eliminate delays in payments.*
3. *Business objective further to the prediction – Cost savings beneficial by reducing the number of calls to paying customers with a high likelihood of timely payment.*

The approaches to 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.

Chapter 5: Project Methodology

CRISP-DM framework has been used for this project.

Cross-industry standard process for data mining, knowns as CRISP-DM is an open standard process model that describes common approaches used by data mining experts. It is a widely used analytics model (Wikipedia, 2020).

CRISP-DM breaks the process of data mining into six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The sequences of phases are not strict and move back and forth between different phases as is always required. The arrows in the process diagram indicate the most important and frequent dependencies between phases. The outer circle in the diagram symbolizes the cyclic nature of data mining itself. A data mining process continues after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions and subsequent data mining processes will benefit from the experiences of the previous one (Wikipedia, 2020).

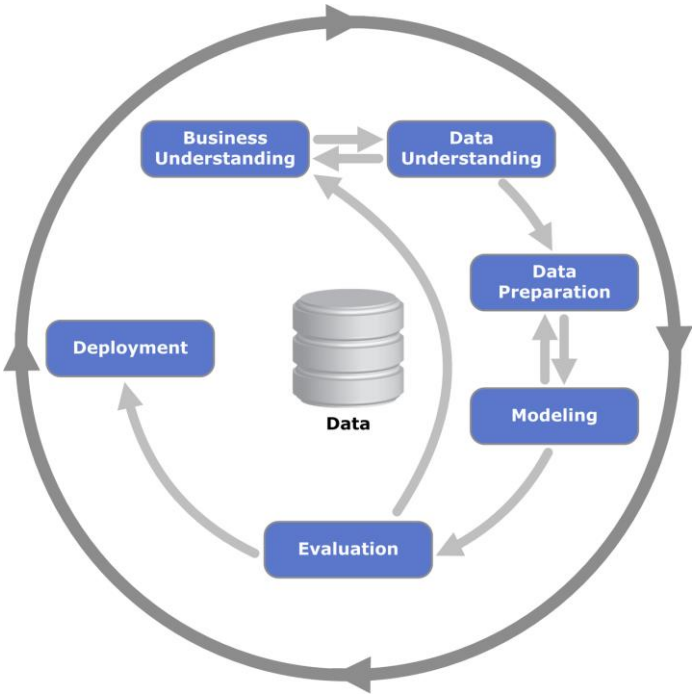


Figure 5.1 “CRISP-DM Framework” (Wikipedia 2020)

Chapter 6: Business Understanding

Cash Flow is a financial metric that measures, as its name implies, the flow of funds within an organization. It is the difference between the money that enters the company (cash inflow) and the money that leaves the company (cash outflow). Profit is another commonly used metric, which is calculated by subtracting revenues from expenses indicating whether the company is running at a profit or a loss.

Money inflows and outflows can be of different categories according to their operating, financial and investment cycles (Lopes & Rebelo, 2021).

In table 6.1, some examples are presented (Rebelo, 2022)

	Cash inflows examples	***Cash outflows examples***
Operations	Receiving a payment from a good or service to a customer	Paying a good or service to a supplier
Investment	Proceeds from disposal of a property or equipment	Payments for a property rent or equipment sold (longer than operating cycle)
Financing	Contracting debt	Debt liquidation

Table 6.1 Examples of Cash Inflows and Cash Outflows

In an operations cycle, for example, tools or services might be purchased from a supplier (outflow) to perform machinery maintenance for a factory or a product solution, and the customer may pay for this service (inflow). Considering the operation cycle, one can estimate the Operating Cash Flow, which results only from operating activities, and the same applies to the investment and financing categories (Lopes & Rebelo, 2021).

There is a consensus that improving the Order-to-Cash (O2C) process will help to improve cash flow management as well. The O2C process presented in Figure 6.1 includes all the steps that take place at the company level, from when a potential client is assessed for credit to when a payment is received. These steps

are common to most companies, and usually and the invoice information is stored in ERP systems (Korotina et al., 2015).

The account receivables management can be in a sub-process inside the O2C, Invoice-to-cash (I2C) (Rebelo, 2022). This process will start once the invoice is created and sent to the customer and its duration depend on the collection effectiveness (Zeng et al., 2008).

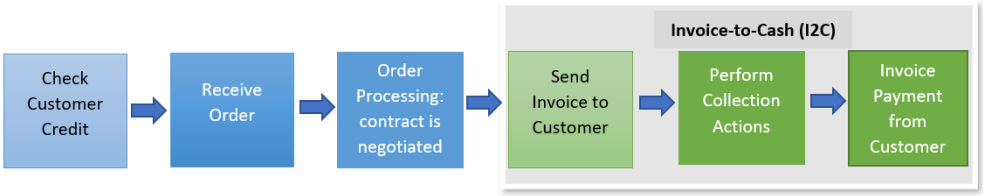


Figure 6.1 The Order-to-Cash (O2C) and Invoice-to-Cash (I2C) processes based on (Zeng et al., 2008)

This is where the AR come in, which is the object of this study. In accounting, AR refers to the cost to be paid to a vendor for the product or services it has supplied or implemented and has not been paid by a consumer. Customers owe AR for purchases made on credit (Investopedia, 2021). Customers are given credit by companies to make payments some days, weeks or even months after the invoice is issued. As a result of this lagging time, the operation cycle length of the organization is determined (Quiry et al., 2009). Contracts between the two entities are always different and vary according to a variety of factors. A higher credit rating of a customer allows it to negotiate better credit terms and contracts, such as lower capital rates and discounts (Pfohl & Gomm, 2009). As a result, cash flow inflows from accounts receivables usually arrive later, and sometimes they may not even arrive on time, affecting the calculation of cash flow. Managing the receivables correctly is an essential step in estimating cash flow in any industry, so monitoring receivables and lagging times is an essential step (Rebelo, 2022).

Some of the critical concepts in collection management are reviewed: in a transaction contract, the **payment term** is defined as days provided for an

invoice transaction cost being paid (from invoice created date to due date), usually, 30, 45, or 60 days and the **due date** is the maximum time when the transaction cost must be paid. An **overdue** (**or late**) invoice is a bill that is late in its payment, i.e., the invoice's due date is passed, and the customer has not yet made the payment whereas, in contrast, **outstanding invoices** are not paid yet but are not late either, their due dates lie in future. Invoice can also be **paid in advance**, earlier than the agreed due date. The customer is typically at risk when making these advance payments as they can pay for the service or good, but not receive it at the end (Pfohl & Gomm, 2009).

The client is a large MNC that sells software products and services in business applications and consulting.

This project aims to provide algorithmic solutions to the team in predicting the delay in invoice payments based on the features like reasons for which the invoice is created, payment amount, type of invoices whether e-invoicing or web, Purchase Order Mandate, and the Country in which the customer is located. Based on the proactive predictions, delays in invoice payments can be reduced.

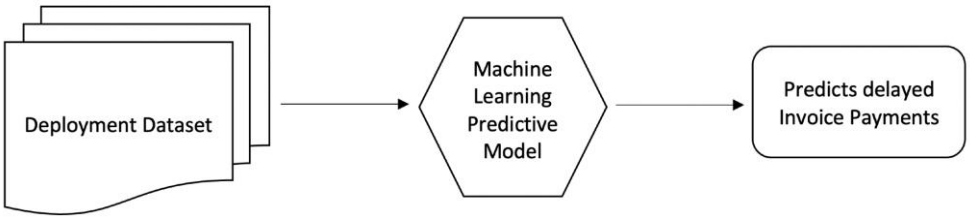


Figure 6.2 Algorithmic Solution for Delayed Invoice Payment

Figure 6.2 depicts the solution for the delayed invoice payments. Dataset is sourced to supervised machine learning model which predicts delayed invoice payments along with the probability of delay in weeks.

In this dataset, we have more than 50% defaulting the payment. With historical examples of invoices and their end result, the project aims to answer the following questions for a new invoice:

- Will the invoices be paid late or on time?
- What could be the duration in which the invoice will be paid?

An attempt is being made to develop a classification system to identify the category a new invoice belongs to, based on a training set of data that includes cases with known outcomes.

As a supervised learning problem, formulate the invoice outcome prediction problem given the consequence of the historical invoices, a model is created for forecasting the payment time in days for a new invoice.

Chapter 7: Data Understanding

7.1 Collecting the Initial data

The analysis in the following sections is based on a fortune 500 company's invoice data set that provides technology services. The invoice data is in quarterly frequency and have collected data from January 2018 Jan to September 2020.

Data collected from Organization Reports are structured and in masked format. Data is masked so that the organizations and customers' confidentiality can be maintained.

The Table 7.1 consists of the list of features in the given 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.
DPLC_Week (weeks)	Days paid late converted to weeks.

Table 7.1 Data Dictionary

FY_Quarter	Quarter	Transaction Numb	Transaction Type	Transaction Da	Concat	Batch Source_I	Transaction Da	Transaction Date_Mo	Transaction Date_D	Due Date	Due_Date_Date
Q1FY19	Q1	Trx_1	trx_type-1	21-Mar-17	Trx_142815	batch_source-1	21	Mar	Tue	5-May-17	05
Q1FY19	Q1	Trx_2	trx_type-2	24-Dec-13	Trx_241632	batch_source-2	24	Dec	Tue	8-Jan-14	08
Q1FY19	Q1	Trx_3	trx_type-3	31-May-16	Trx_342521	batch_source-2	31	May	Tue	30-Jun-16	30
Q1FY19	Q1	Trx_4	trx_type-4	8-Dec-16	Trx_442712	batch_source-3	08	Dec	Thu	23-Dec-16	23
Q1FY19	Q1	Trx_5	trx_type-3	23-Feb-17	Trx_542789	batch_source-2	23	Feb	Thu	25-Mar-17	25
Q1FY19	Q1	Trx_6	trx_type-3	14-Mar-17	Trx_642808	batch_source-2	14	Mar	Tue	13-Apr-17	13
Q1FY19	Q1	Trx_7	trx_type-2	22-Apr-17	Trx_742847	batch_source-2	22	Apr	Sat	22-May-17	22
Q1FY19	Q1	Trx_8	trx_type-5	27-Apr-17	Trx_842852	batch_source-4	27	Apr	Thu	12-May-17	12
Q1FY19	Q1	Trx_9	trx_type-2	31-May-17	Trx_942886	batch_source-4	31	May	Wed	30-Jun-17	30
Q1FY19	Q1	Trx_10	trx_type-2	31-May-17	Trx_1042886	batch_source-4	31	May	Wed	30-Jun-17	30
Q1FY19	Q1	Trx_11	trx_type-4	8-Jun-17	Trx_1142894	batch_source-2	08	Jun	Thu	8-Jul-17	08
Q1FY19	Q1	Trx_12	trx_type-2	20-Jun-17	Trx_1242906	batch_source-2	20	Jun	Tue	20-Jul-17	20
Q1FY19	Q1	Trx_13	trx_type-4	28-Jul-17	Trx_1342944	batch_source-2	28	Jul	Fri	27-Aug-17	27

Due_Date_Date	Due_Date_Mor	Due_Date_D	ABE	Credit Hold	PO Mandate	Web Invoicing	E Invoicing	DPLC	USD_AMT	Paid_15	buffer_da	DPLC_We
05	May	Fri	No	No	No	No	Yes	459	166	No	45	10
08	Jan	Wed	No	No	No	No	Yes	1668	39,037	No	15	10
30	Jun	Thu	No	Yes	No	No	Yes	733	8,772	No	30	10
23	Dec	Fri	No	Yes	No	No	Yes	573	1,783	No	15	10
25	Mar	Sat	No	No	No	No	Yes	500	3,559	No	30	10
13	Apr	Thu	No	No	No	No	Yes	470	7,661	No	30	10
22	May	Mon	No	No	No	No	Yes	410	22,488	No	30	10
12	May	Fri	Yes	Yes	No	No	Yes	472	16,330	No	15	10
30	Jun	Fri	No	No	No	No	No	360	68,959	No	30	10
30	Jun	Fri	No	No	No	No	No	364	17,289	No	30	10
08	Jul	Sat	No	No	No	No	No	391	7,836	No	30	10
20	Jul	Thu	No	No	No	No	Yes	337	17,985	No	30	10
27	Aug	Sun	No	Yes	No	No	Yes	290	16,369	No	30	10

Figure 7.1 Sample Data in masked format

Sample data in masked format has been depicted in the Figure 7.1.

7.2 Describing the data

The dataset contained 34,752 invoice data, which processes around 1500 invoice transaction data in a month. For confidentiality, customer wise details were not available for study and hence only invoice level details were considered.

Most fields in the dataset were categorical features which posed challenge in terms of variables with no numerical relationship between levels. The most important numerical feature is **USD_AMT**: the value of the transaction in USD. The cost of an invoice transaction was between <1 dollar to ~20M$. There are also some indicative dates such as **Transaction Date** when the invoice was created and **Due Date,** calculated as per the terms of invoice payment.

Few categorical variables in the data are -

ABE (Accelerated Business Expense) - It falls under the definition of accelerated depreciation if it is applied as a method of depreciation for

accounting for income tax purposes in order to allow for greater deductions in the early years of an asset's life (Investopedia, 2021). Figure 7.2 showcases the categorization of ABE for each invoice record.

Count	ABE
No	33941
Yes	811
Grand Total	**34752**

Figure 7.2 ABE

PO Mandate (Purchase Order) – The purchase order is a commercial document that is issued by a company's purchasing department in order to place an order with its vendors or suppliers. Figure 7.3 showcases the categorization of PO Mandate for each invoice record.

Count	PO Mandate
No	34569
Yes	183
Grand Total	**34752**

Figure 7.3 PO Mandate

Credit Hold – By placing a credit hold on a customer's account, the organization is able to prevent them from making further purchases of credit if they have been constantly delaying cost to be paid, having outdo the limits of a credit are categorized as risky. Figure 7.4 showcases the count of invoices breakdown indicating whether the customer has credit hold or not.

Count	Credit Hold
No	32477
Yes	2275
Grand Total	**34752**

Figure 7.4 Credit Hold

Web Invoicing – Digital invoices are typically PDF or Word documents that have been scanned from paper invoices. Figure 7.5 provides an idea of the number of customers opting for web invoicing indicating that more customers are interested in an e-invoice rather than digital invoice.

Count	Web Invoicing
No	33932
Yes	820
Grand Total	**34752**

Figure 7.5 Web Invoicing

E Invoicing – Invoice documents are transferred electronically in the middle of suppliers, buyers. Figure 7.6 showcases that except for a very few exceptions, all the customers prefer e-invoices as it is structured and can eliminate manual ways of sharing the document with the buyer thereby significantly reducing the value and time is being saved.

Count	E Invoicing
No	83
Yes	34669
Grand Total	**34752**

Figure 7.6 E Invoicing

7.3 Target Setting

The main goal of the machine learning models should be able to predict the invoice's outcome. If a new invoice is registered in the system, the model should predict, based on certain characteristics of the invoice, whether there is a timely payment of an invoice cost or late payment (delay, not delayed). If the payment is late, how much delay will the customer incur with the invoice.

The main rationale behind this paper is to classify the outcome of an invoice and make a classification task for the algorithm (Rebelo, 2022).

Paid_15 – This feature, as a binary outcome case, was evaluated whether the invoice will be late or not. Figure 7.7 showcases the measure of this feature, indicating over 47% of invoices defaulted in the payment.

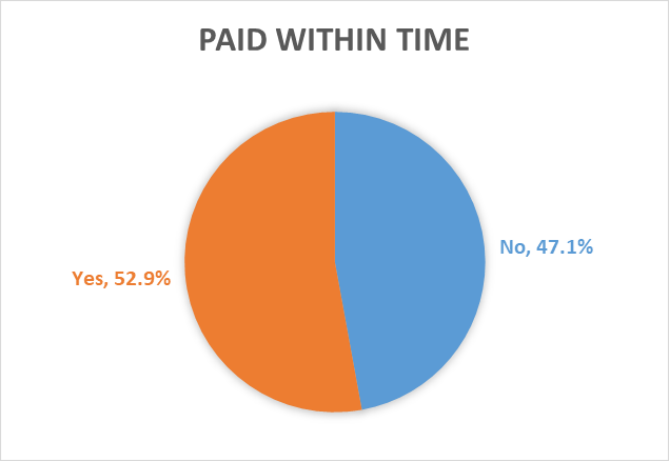


Figure 7.7 Classification of invoices in terms of delay

Literature and authentic experiences suggest that on-time payments can be viewed in various ways. Some consider it as those payments paid on the date of the due date (Zeng et al., 2008), but others consider it a window of days after which are still considered as on-time and are "forgiven" to the customer (Appel et al., 2020b). In this project, an on-time payment was considered with a window of 15 days from the due date. Without data to support a different approach for different customers, this default was established, and standardization was considered an important best practice.

DPLC (Days paid late) – This feature provides the number of delays in days for the invoice which is classified as "Delay = Yes". Figure 7.8 classifies invoices delayed in days into various buckets. Seeing on the graph, 50% of the delays are contributed from the buckets 16-30 and 31-45 days together. We need to pay attention to these invoices and try to find there are common features shared by these invoices.

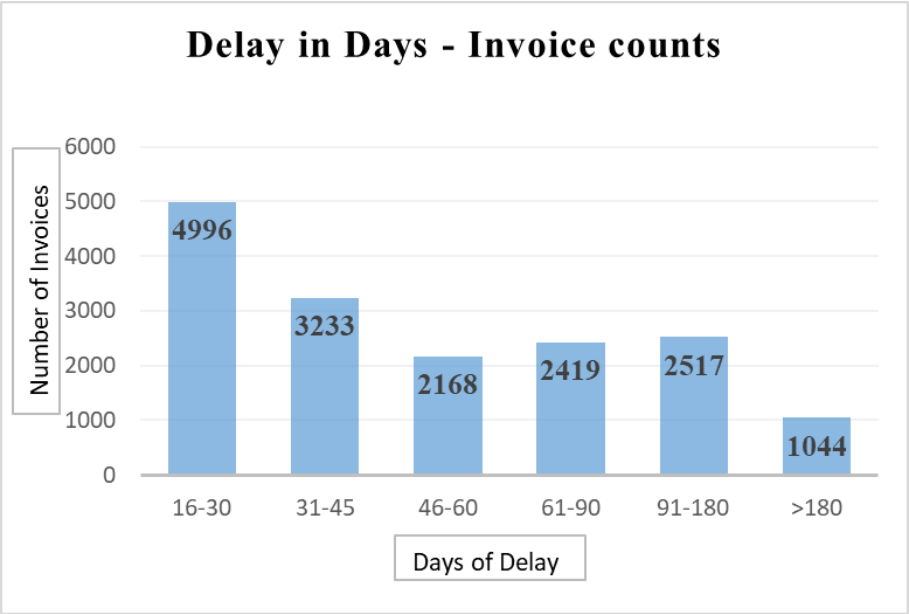


Figure 7.8 Delay buckets in Days

7.4 Exploratory Data Analysis (EDA)

Delay in Invoices – Collections Management Team is responsible for payment collections against open invoices. Challenge is to proactively identify invoices that will be delayed for payment. Currently, the categorization is solely dependent on the past payment behavior of the customers, the value of invoices, and the past payment history of the customers. In a situation where the consumer cannot pay, competing interest within the consumer and vendor is apparent. Therefore, low administration of account receivables may result in a significant percentage of the consumer not paying in between the anticipated time resulting in finances being adversely affected.

An effective way to deal with delays is by reaching out to the customers prior to significant delays. Nevertheless, there should be a fine balance between the intervention with over intervention. To make a decision on whom to contact, it is ideal to know upfront the list of customers going to delay payment.

The object of the dataset used in this project is the Account Receivables from a Fortune 500 Company. Out of the dataset considered (34,752), **53%** of invoices

had a positive outcome, meaning these invoices were closed without delays and **47%** had delays in payment.

Figure 7.9 depicts the first level of categorization if the invoice will have delays or not. Out of the overall dataset, 53% of the invoices were paid on time and 47% are delayed invoices.

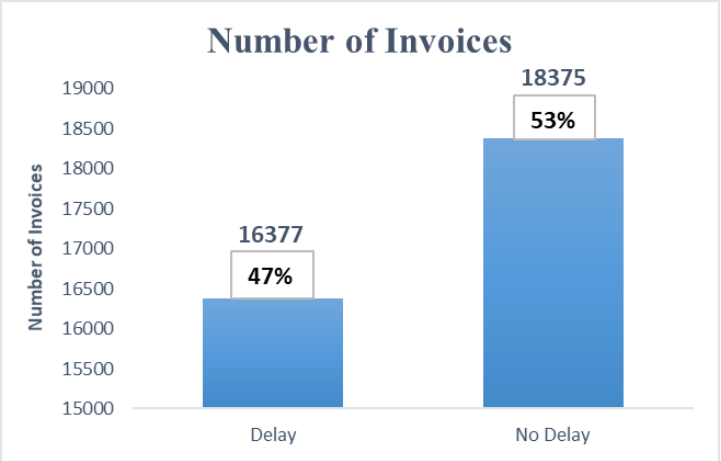


Figure 7.9 Invoice Delay vs No Delay

The payment classification was divided into 2 classes according to the level of delay of an invoice. The invoices can be classified into -1 for early payments, 0 for on-time payments.

Based on account receivables management best practices, the delayed are in turn classified into late (1) if the payment fall between 1 to 30 days after the due date; vary late (2) if the payment fall late between 31 and 90 days after the due date; and critically late (3) if the payment falls late 91 days and onwards. It is important for the vendor to identify those customers likely to pay earlier than the due date on the contract to define collection strategies.

Out of 53% of invoices that had no delays in payment, over **47%** of invoices out of these positive outcomes were **paid earlier** than the due date and **6**% were paid **on time**. It is depicted in Figure 7.10 showcasing the distributions of the 2 constructed targets, by number of invoices.

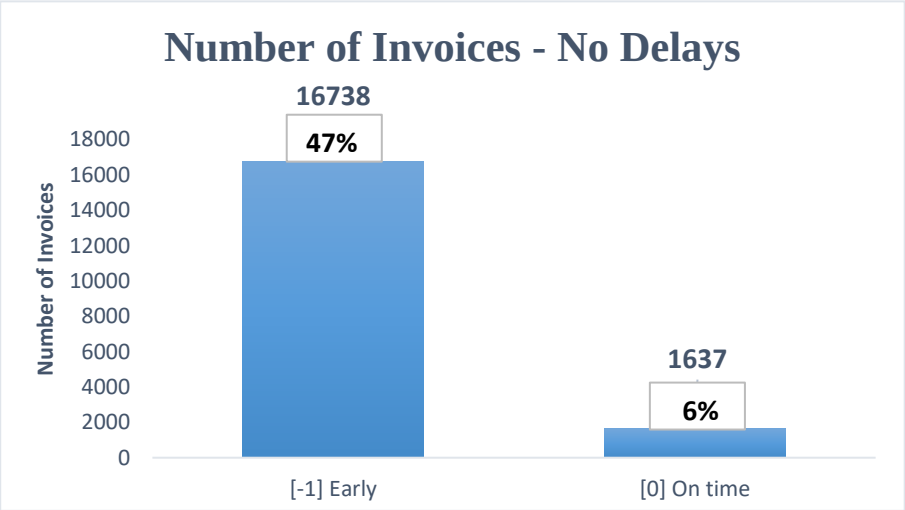


Figure 7.10 Invoice Categorization – No Delay Payments

Out of **47%** of invoices were delayed for payments, **29%** were up to 1 month late, **11%** were between 1 to 3 months late, and **7%** were delayed for more than 3 months. Depicted in Figure 7.11 are the distributions of the three constructed targets, by number of invoices. Invoices that are delayed for much longer are the most problematic and therefore called "Critically late" invoices.

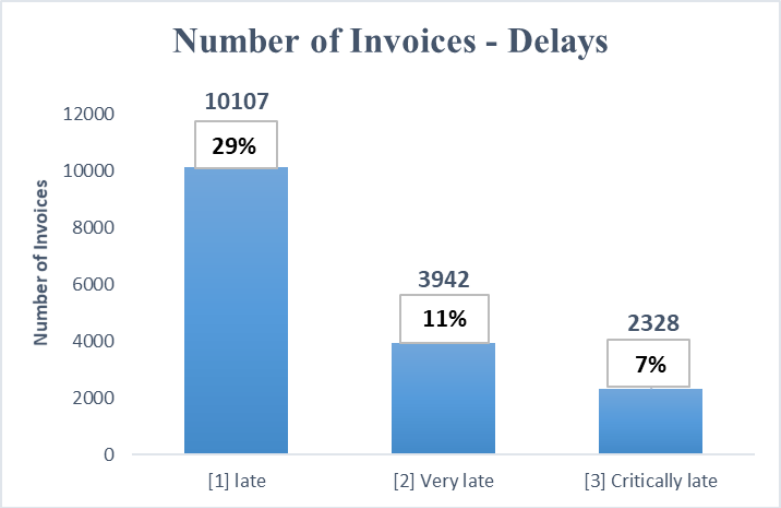


Figure 7.11 Invoice Categorization – Delay Payments

The requirement is that every invoice that is open needs to be paid based on the agreement the company has with the customers while booking orders. The invoice payment must be done within 30 days from the transaction date. The

Collections Team starts following up on the payments during this period. But a certain set of invoices or customers or LOBs, invoices may get delayed and go beyond due dates. There is a grace period of 15 days provided for the payment, and hence there is a 45-day term provided for a customer.

There can be valid delays from the organization's end such as the product is not delivered on time, movement to the cloud, and various other factors. Big customers have multiple products configured for them. Specific factors such as time of the year, LOB, Order Type can yield to payment delays. There can also be delays with respect to e-invoice or a physical copy of the invoice to be delivered to the customer.

Figure 7.12 shows the total invoices information for the entire dataset and the number of delays. The primary objective of the project is to detect the potential delayed invoices. The invoices have been grouped into two classes – invoices paid on time and not paid on time. In figure 7.9 we could see that over **47%** of times there have been delays in payments.

Time	# of invoices	#of delayed invoices	% of delayed invoices
2018-Qtr1	252	252	100.0%
2018-Qtr2	3387	2181	64.4%
2018-Qtr3	3703	1637	44.2%
2018-Qtr4	3344	1467	43.9%
2019-Qtr1	3420	1447	42.3%
2019-Qtr2	5702	2569	45.1%
2019-Qtr3	3554	1579	44.4%
2019-Qtr4	3063	1396	45.6%
2019-Qtr1	3450	1452	42.1%
2019-Qtr2	3526	2008	56.9%
2019-Qtr3	1221	260	21.3%

Figure 7.12 Delayed Invoice Summary

The number of invoices created over the past few years has remained stable as shown in Figure 7.13, 7.14 and 7.15 for years 2018-2020. There has been slight upward trend in 2018 and 2019 and December is usually the weakest in terms of invoices created, possibly due to holiday season.

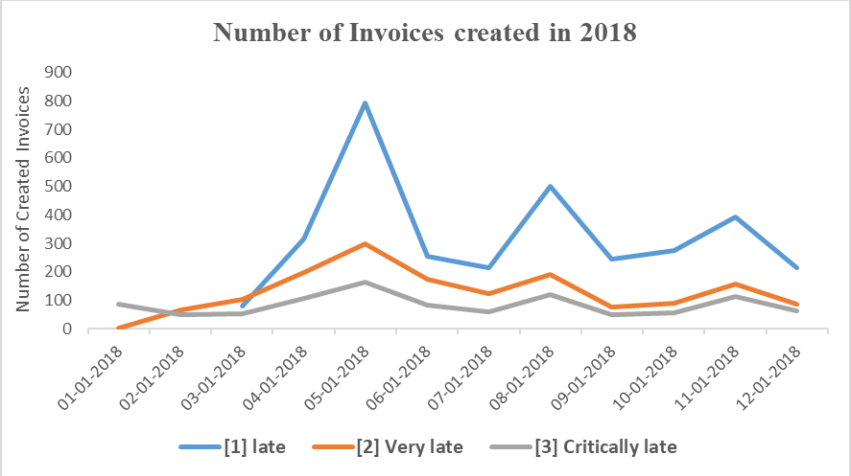


Figure 7.13 Number of Invoices created in 2018

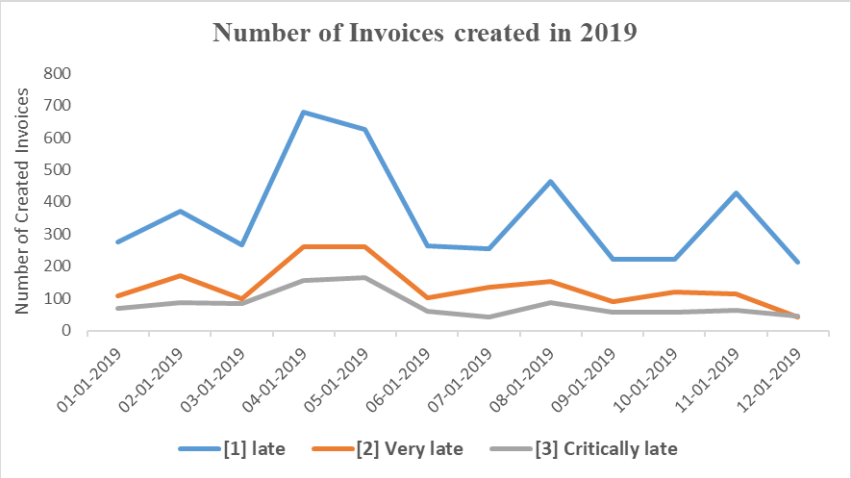


Figure 7.14 Number of Invoices created in 2019

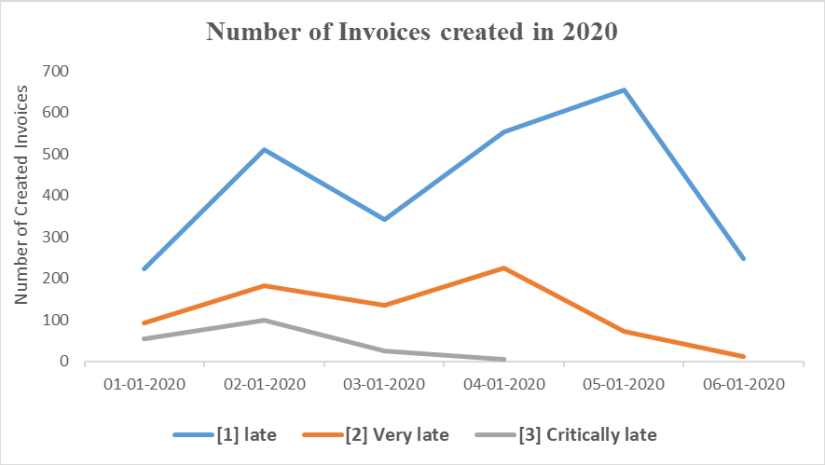


Figure 7.15 Number of Invoices created in 2020

We can see a steady increase in total invoices created between March to May 2020 as shown in Figure 7.15 even though these months correspond to COVID-19 lockdown that were established in many countries. Further data analysis for the last 2 years can be a good study to understand the pattens during pandemic.

Payment terms: Invoices also include the terms of payment (Rebelo, 2022). Invoice payments are subjected to a "buffer" period after issuing an invoice. In the Figure 7.16, payment term for this dataset considered in the study is shown. From Figure 7.16 it is evident that the yardstick for payment term is 30 days for most of the invoices.

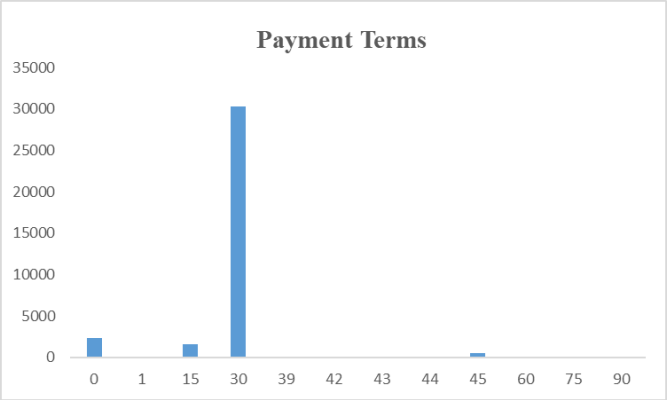


Figure 7.16 Sample Payment Term

Invoice amount and delay: It may be naive to assume that the reason for invoice delays is due to the amount of the invoice: the higher the amount of the invoice, the more likely the delay. In order to ensure the solidity of finance, is there a likelihood that clients can delay the invoice payment if the amount of sales order is higher?

To answer that question, a box plot can be considered as shown in Figure 7.17 for analysis of late versus on-time payments. As mentioned earlier in the previous section, for the invoice, two end results can be seen. The first consequence is checking whether there is a delay or not as plotted in Figure 7.17. The cases in terms of late, very late, or extremely late classes are plotted in figure 7.18 which is asking how the delay would be.

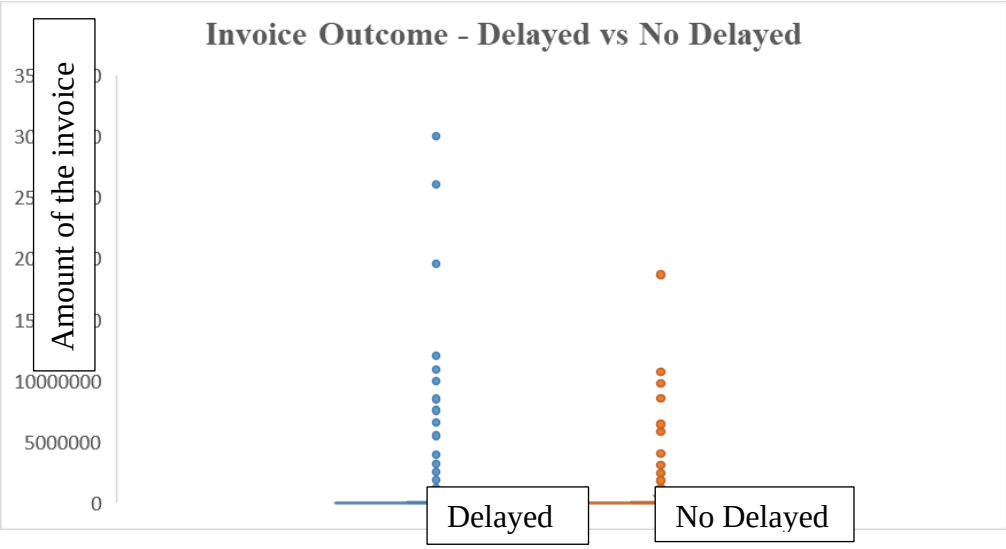


Figure 7.17 Invoice amount - delay or not

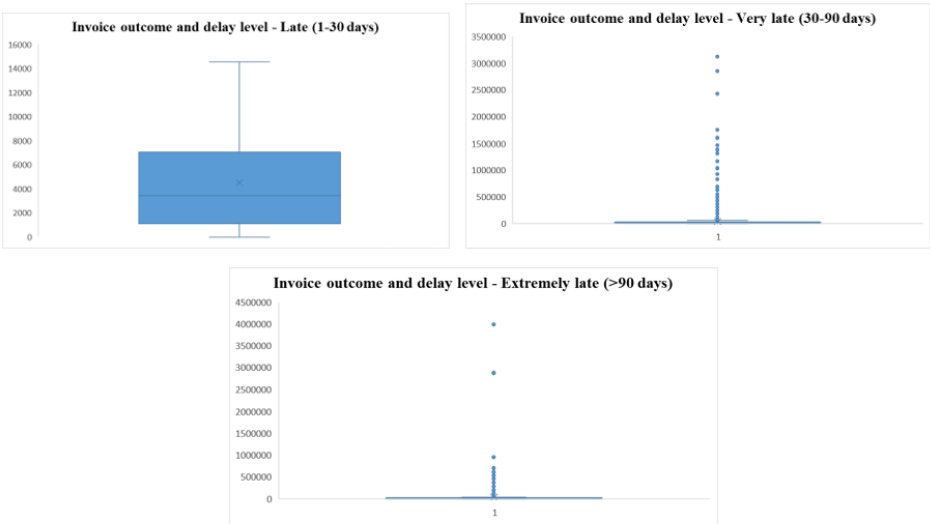


Figure 7.18 Invoice amount and delay level

Further to Figure 7.17 and Figure 7.18, the next analysis plots the “Days Late vs Amount” in Figure 7.19. The amount of an invoice does not appear to be correlated with the late payment of invoices.

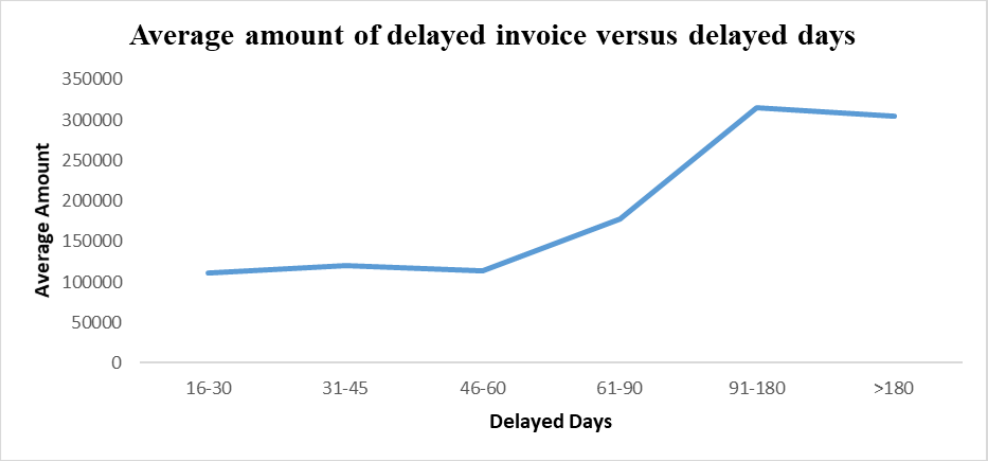


Figure 7.19 Days late versus Amount

Additionally, we are reminded that it is difficult to predict the delay in an invoice by using just one factor. In order to understand and predict the invoice, more information will be collected about the invoice and will use more advanced models.

There is a need to collect more data on this to understand the delay in payments and work on reducing the delays.

Chapter 8: Data Preparation

Data preparation is a step that requires the most time and effort. Data Preparation stage aims to address different problems as mentioned below

Selecting columns and dropping duplicate rows
Treating Missing Values
Checking for Outliers
Coherence Checking
Data Transformation and Feature Engineering

Data is collected from the Database Reports i.e., downloaded in the .csv format. Data is then masked and shared for analysis and study as showcased in Figure 8.1.

Figure 8.1 Data Extraction Flow

The time window considered for the study is January 2018 to September 2020. There are no missing data in any of the columns and hence no further imputing is required.

Based on the Exploratory Data Analysis (EDA) and the domain experts' suggestions, the following features have been removed for further analysis as shown in Table 8.1.

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.
DPLC_Week (weeks)	Post factor feature.

Table 8.1 Reasons for dropping features

Data Preparation Steps:

- Dataset is split into three categories training, validation and test set.
- Training and validation dataset contains Paid_15 status as Yes or No indicating if there is a delay in payment.
- Test set is used as the deployment data for re-validation of the model.
- One hot encoding has been done on the categorical data.

8.1 Selecting data and dropping duplicate rows

Out of 23 columns, only 14 have been selected candidates for feature engineering part. It was based on several characteristics of each column. In short, irrelevant columns that do not represent valuable information to the problem were dropped. This is the case for columns such as FY_Quarter, Quarter, Concat, Batch Source_Masked.

Categorical columns with too many levels are also irrelevant. In fact, their use will undermine the model's performance as there will be fewer examples available for each level of training, which will reduce the ability of each level to have any impact on the model's outcome (Quiry et al., 2009).

It was checked and found that there were no **duplicate rows** on the rows level in the dataset.

8.2 Treating Missing Values

Missing values refer to data points that are not present in the dataset. They could be missing at random, for example, because the data was entered incorrectly, but their absence could also be indicative of something so that their absence is not random (Kang, 2013).

There are no missing data in any of the columns and hence no further imputing is required.

8.3 Checking for Outliers

Outliers are values that differ significantly from the rest of the data. It is possible that they represent errors that were accidentally inserted into the dataset, or they may simply represent extreme values that are demonstrating variance (W et al., 2018). Domain knowledge can be used to differentiate between the two.

Invoices for which payment is delayed or advanced over the course of one year are rare and usually involve special contracts that should be considered separately. Furthermore, all payment terms that are longer than 1 year (the period between the transaction date and due date) have been deleted from the system.

8.4 Feature Extraction:

The last step of the Data Preparation stage is feature engineering. In machine learning, models learn by searching for patterns in their input data. That input is what practitioners call the features: an informative representation of the data in numerical form (Zeng et al., 2008).
During the feature engineering phase, data is transformed in order to create predictive features for the model to interpret.

The Data Understanding component outlined the details of all the attributes considered in this dataset. Following are some of the issues with respect to attributes:

- It is possible that the details of attributes are enormous and dispensable for management.
- There have been scenarios in which categorical data is saved numerically.
- Customer information is not sufficient but can be gathered.

By using machine learning to predict the late payment of invoices, a subgroup of attributes has been chosen to create a new set of attributes in order to promote discoveries as well as to boost the interpretation of the results.

New features have been created to train the classifier. A delayed invoice transaction payment may be the result of multiple factors. There is a direct correlation between the time spent in processing an invoice and its features along with other invoices that are simultaneously processed. If the volume of invoices is higher than what can be handled at a given point in time, this may result in congestion at a few points in the process. Lesser invoices can increase the time spent on processing. Feature extraction and selection become highly crucial in increasing the accuracy of the model.

8.4.1 Choosing the right data

To process invoice data, the initial footstep is choosing the right attributes from the dataset. This would correspond to questioning the kind of particulars, if any, suitable to relate to the cost when an invoice is considered. Table 8.2 consists of attributes listed below.

Table 8.2 below mentions the cost and dates of invoices. Due to privacy issues, it is not able to gather information on products or services and customer information.

Name	Meaning
Transaction Number	Number of each transaction
Transaction Date	Date of the transaction created
Due Date	The latest date to receive payment of an invoice
USD AMT	Net cost of the invoice
Payment Term	Agreed terms of payment
Credit Representative	Whether credit provided or not

Table 8.2 Key attributes of the data

8.4.2 Two levels of features

The attributes, as illustrated in Table 8.2, furnish meagre facts, especially regarding the client associated with a particular transaction. However, it is

necessary to know more about the virtue of the payer in order to forecast the settlement of the individual transaction.

An invoice has two grades of attributes:

- Transaction grade
- Buyer grade.

A transaction's grade attributes include the cost, settlement, and different transaction dates. In our study, there was only a transaction, and the due date is available. The project aggregates the historical invoices at the invoice level not at the customer level due to confidentiality and privacy of data.

The data contains several important components, mentioned below:

1. How many invoices have been paid?
2. How many late invoices?
3. Cost of invoices that have been paid.
4. Late invoices cost
5. Ratio of Lag (Between 1 & 2)
6. Ratio of lag of cost (Between 3 & 4)
7. Average settlement
8. Number of days delayed averaged out

8.4.3 Extra information and unexpected features

Extraction or selection of attributes has been regarded as an amalgamation of skill and accuracy in the machine learning section (Rebelo, 2022). Considering that the algorithm has been intended to ascertain patterns of cost of an invoice transaction, additional intelligence on an invoice transaction be feasible behind the cost of an invoice transaction along with the monetary steadiness. Stable cash flows are an important element of monetary steadiness, mainly towards a month’s end, during a time where usually salaries and remaining payments are

paid. In case an invoice transaction is expected to be paid at a month's end, this may increase the chance of it being late.

8.4.4 New features extracted for learning

New features have been created to increase the efficiency and performance of machine learning models. Table 8.3 details the features created in this study.

Features	Explanation
Delay	Transaction Date – Due Date; Delay in Days
Due_Date_Date	Date part from Due Date
Due_Date_Month	Month part from Due Date
Due_Date_Day	Day part from Due Date
Transaction_Date_Date	Date part from Transaction Date
Transaction_Date_Month	Month part from Transaction Date
Transaction_Date_Day	Day part from Transaction Date
Average_Delay	Average delay of the delayed invoices
Ave_Buffer	Average payment term
Delay_Bucket (Days)	total delay in days between Transaction Date and Due Date bucketed for categorization of invoices
Delay_bucket_id	ID created for delays

Table 8.3 List of New Features

Label encoding and one-hot encoding – A typical data collection comprises a lot of attributes and a culmination of quantitative and qualitative attributes. One-Hot encoding was done to the Paid_15 label to create new attributes on basis of distinctive features in the attribute set.

Chapter 9: Modeling

In this section, an overview of machine learning procedures is provided and explained the process of predictive modeling.

Data discussed in the previous section has been fed into multiple models to get the predicted values of disputed invoices. Based on the problem statement and the data availability, the Classification algorithm, a supervised learning technique has been considered which is used to predict the categorical observations based on the training set. Programs learn from the given data and classify it into classes or groups.

Before implementing the machine learning models, data sets are arbitrarily divided into training and testing sets. 80% of invoice data is used as training sets and 20% of data is used as testing sets. The training set is used to train the classification model with certain features of invoices, such as the amount of the transaction. A Testing set is used to evaluate the performance of classifiers.

9.1 Evaluation metrics

Before the algorithms are introduced to tackle the classification problem, there is a need to clearly define the metrics used for evaluating them. In this study, binary classification is applied to classify on-time or late payments and regression to predict the delay in the number of days for payment.

A confusion matrix was used in this project to compare the number of predicted instances in each class to the correct classification. The confusion matrix for the binary problem can be described below in Table 9.1.

		Predicted Classes	
		On-time Payment (Positive Class)	Late Payment (Negative Class)
Actual Classes	On-time Payment (Positive class)	True Positive (TP)	False Negative (FN)
	Late Payment (Negative Class)	False Positive (FP)	True Negative (TN)

Table 9.1 Confusion matrix metrics for binary classification of invoices

In case of false positives, the organization will expect the customer to pay earlier than he will be paying and not take any action to prevent the late payment behavior. The company could lose financial benefits if the money is received from that account later than expected instead of putting it in the bank sooner than expected. However, the organization will also attempt to target customers who have exhibited bad behavior with more calls to action and stricter terms and conditions. In this manner, the wrong (well-behaved) customers are targeted as such, and such unnecessary actions will eventually result in relationship degradation and a decrease in customer satisfaction. Hence the true objective to measure the algorithm performance is how well the model can accurately predict the delay.

Accuracy – Accuracy score is the percentage of correctly classified invoices as on-time or late. Since it can be easily comprehended by practitioners and cash collectors alike, accuracy is a valuable metric for this problem.

Precision – It is the count of positive observations divided by all positive class values.

Recall – It is the proportion of correctly predicted on-time payments, over the sum of all actual on-time payments either correctly predicted or not. In this case, the model must be able to accurately identify the positive instances.

F1 Score – F1 Score conveys the balance between precision and recall.

The **Receiving Operating Characteristics Curve** (ROC) and the **Area Under the Curve** (AUC) have been considered to assess the predicting ability of the model.

9.2 Machine Learning Algorithms for Supervised learning

Following classification algorithms were applied for this dataset:

- XGBoost Classification
- XGBoost Regression

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

Predicted Features –

Paid_15: Classify in case of invoice transaction cost will be paid in 15 days after due_date or not.

DPLC: Predict the delay in payment in weeks.

XGBoost (Extreme Gradient Boosting) Classification: XGBoost is an ensemble method combining multiple models rather than using a single one is an effective method of increasing the accuracy and strength of outcomes. This requires numerical features; categorical features were encoded using One-Hot encoding in the model built in this study. It does not accept missing values. Additionally, to achieve good performances, interaction variables are necessary for accurate predictions, and it is recommended that inputs should be linearly scaled (W et al., 2018). One important assumption is that there should not be any multicollinearity present in the independent variables (or highly correlated features), otherwise there is a real chance of overfitting (Brownlee, 2016)

In this study, XGBoost Classification was used for predicting whether there is a delay in payment or not. The accuracy achieved is **81.1%** and F1 Score is **0.66**.

XGBoost Regression: XGBoost Regression was used to predict the delays in weeks for those invoices that are predicted as delays in the Classification model. Those invoices having delays as “Yes” are provided as input to the regression model which has predicted the delays in weeks for each invoice. Trial and error can be used to find appropriate hyperparameters for a given dataset, or systematic measures can be taken, such as using a grid search across a range of values. The model in our study predicted an R2 Score of **0.11** and an RMSE of **2.83**.

Classification algorithm - Random Forest (RF): The Random Forest approach combines multiple learning algorithms to achieve better predictability. For the classification problem, the random forest grows an ensemble of trees and makes a prediction based on the majority votes of trees (Breiman, 2001).

In this study, RF Classification was used for predicting whether there is a delay in payment or not. The accuracy achieved is **81.0%** and the F1 Score is **0.65**.

Random Forest Regression: This type of regression technique aims in performing cross-validation in order to improve accuracy. It can work well with missing values along with the steady maintenance of accuracy.

Regression was used to predict the delays in weeks for those invoices that are predicted as delays in the Classification model. The model in our study predicted an R2 Score of **0.16** and RMSE of **2.76**.

Neural Network Classifier: Neural Network (NNet) is inspired by the biological neural network in the human brain, which is usually used for an abundance of data with little underlying theory (Rebelo, 2022).

The algorithm was used for predicting whether there is a delay in payment or not. The accuracy achieved is **32.0%** and the F1 Score is **0.10**. This is the least accuracy and F1 Score achieved amongst all the models considered.

K Neighbors Classifier: K-Nearest Neighbors (KNN) stores the training examples and perform the classification of new data points based on training data points for the closest representation. It is a “lazy” algorithm because it delays the searching until a new data point has to be classified (Kuhn & Johnson, 2013), and it simply does not learn anything rather, it performs a “look-up” on the training data (Rebelo, 2022)

KNN was used for predicting whether there is a delay in payment or not. The accuracy achieved is **45.0%** and the F1 Score is **0.20**.

Linear SVC (Support Vector Classifier): Support Vector Machine (SVM) has been considered in the case of supervised learning problems but is mostly utilized in classification problems. The data item is plotted in the SVM algorithm as a point in n-dimensional space (where n is the number of features), along feature having considered a particular coordinate. Following that, we identify the hyperplane that best distinguishes the two classes.

In this study, SVC was used for predicting whether there is a delay in payment or not. The accuracy achieved is **55.0%** and the F1 Score is **0.30**.

Results

In summary, XG Boost Classifier and Random Forest Classifier have the highest Accuracy and F1-Score as compared to other models.

To improve the probability of prediction of invoices and their classification, it is highly imperative to gather data on customer behavioral patterns, multiple transactions from a single customer, demographical data, and seasonality.

Chapter 10: Model Evaluation

Accuracy of the models of XGBoost Classifier, Random Forest Classifier (RF), Neural Network Classifier, K Neighbors Classifier, and Linear SVC (Support Vector Classification) and R2, RMSE Score for XGBoost Regressor, Random Forest Regressor are showcased in Table 10.1 and Table 10.2 respectively.

Paid label		
Models	**Accuracy**	**F1 Score**
XGBoost Classifier	81.10%	0.66
Random Forest Classifier	80.80%	0.65
Neural Network Classifier	32.80%	0.12
K Neighbors Classifier	45.02%	0.20
Linear Support Vector Classifier	55.02%	0.30

Table 10.1 Metrics for Paid label

DPLC (Delay in Weeks) label		
Models	**R2 Score**	**RMSE Score**
XGBoost Regressor	89.40%	2.87
Random Forest Regressor	92.10%	2.86

Table 10.2 Metrics for DPLC (Delay in weeks) label

Key Metrics –

Let us look into some of the key metrics as an outcome of the model.

Confusion matrix: -

		Prediction	
		On Time	*Delay*
Actual	*On Time*	1005	675
	Delay	530	1266

Figure 10.1 – Confusion Matrix of Invoice Delay Prediction for test result

The Confusion metrics of the sample test are shown in Figure 10.1. The accuracy is **74%.**

Precision: - The Precision score is **0.65** for the model.

Recall – 0.6 – 0.7 is the score.

F1 Score – 0.67 is the score**.**

ROC (Receiver Operating Characteristic) Curve: -

The performance or accuracy of a classifier to distinguish two classes of data is a measure of the ROC curve. Figure 10.2 shows the ROC curve for this study with an accuracy of **73%.**

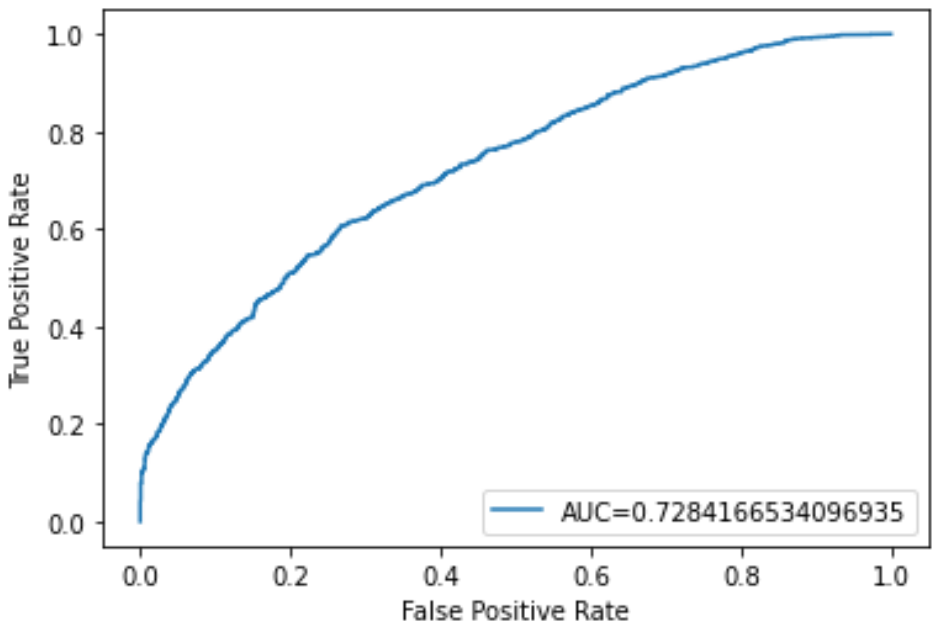


Figure 10.2 ROC Curve

A small set of **1000** invoices have been considered for testing the performance of the model. This data is passed to XGBoost Classification and XGBoost Regression model for classifying whether an invoice will have a delay or not along with the delays in weeks respectively.

XGBoost classifier was able to classify **602** (out of 1000) invoices as delay = "Yes", indicating that 60% of invoices have been classified as delay in payment.

Random Forest Classifier was able to classify **374** (out of 1000) invoices as delay = “Yes”, indicating that 37% of invoices have been classified as delay in payment.

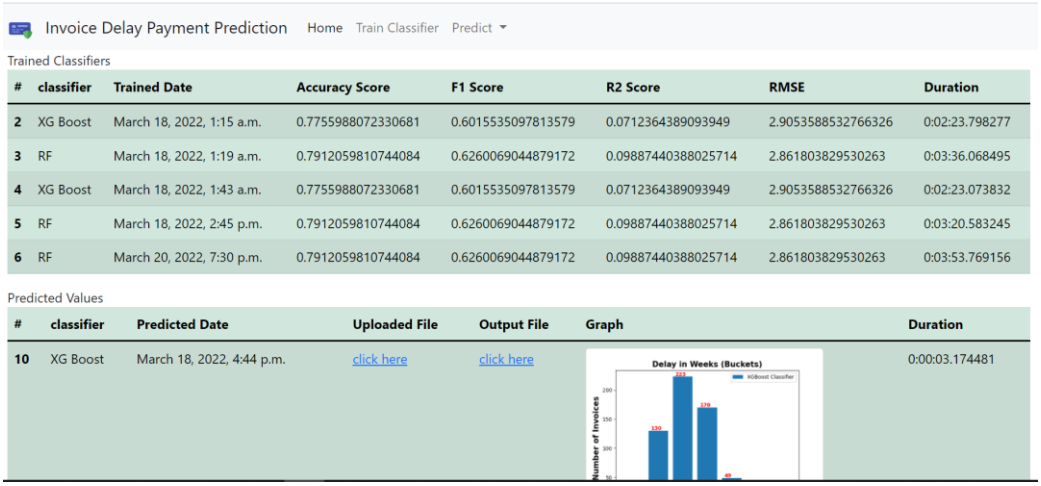
The model prediction should be used to focus on defaulters instead of reaching out to a customer who pays on time assuming the cost per call is 1000$ (350*1000 = 350000), the company can save 40% cost by not contacting those customers who will pay on time.

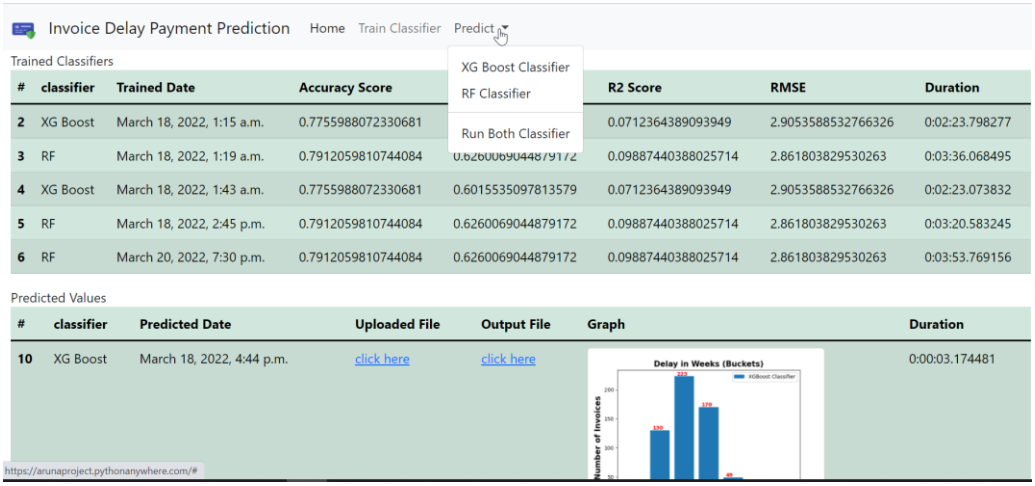
Chapter 11: Deployment

The model will be deployed on a web application which will provide an option for the user to upload an excel file consisting of invoices to be predicted for delays. The output file for XGBoost Classifier and Random Forest Classifier consists of predicted delay in weeks. A new set of invoices can be passed through the model for prediction. This solution shall be provided to the Collections Team for the prediction and proactive follow-ups for delays.

The application has been developed using Django which is Python's framework for the web allowing fast evolution along with practical design.

Link: https://arunaproject.pythonanywhere.com/





Chapter 12: Analysis and Results

Test datasets consisting of 1000 were passed through the models. XGBoost Classifier, XGBoost Regressor, Random Forest Classifier, and Random Forest Regressor models were used for the prediction of Paid label and DPLC (Delay in weeks) label.

XGBoost Classifier was able to predict 602 invoices out of 1000 records will have a delay in payment and XGBoost Regressor was able to provide the number of delays in weeks. For understanding purposes, delay in weeks has been capped to 10 weeks as the maximum period.

Random Forest Classifier was able to predict 374 invoices out of 1000 records will have a delay in payment and Random Forest Regressor was able to provide the number of delays in weeks.

Row Labels	Count of Predicted Delays
No	398
Yes	602
Grand Total	1000

Table 12.1 XGBoost Classifier Prediction

Row Labels	Count of Predicted Delays
No	626
Yes	374
Grand Total	1000

Table 12.2 Random Forest Classifier Prediction

Figure 12.1 is the final output of the XGBoost Regressor for delay in weeks. Weeks have been considered in buckets and any delay beyond 10 weeks has been capped to the limit of 10 weeks.

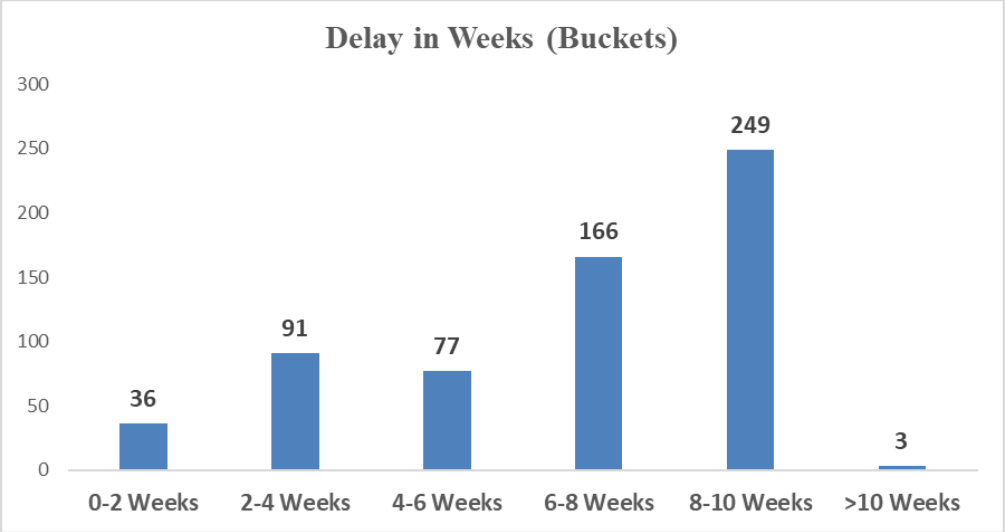


Figure 12.1 Delay in weeks buckets for XGBoost Regressor Output invoices

Chapter 13: Conclusions and Recommendations for future work

This project solely focuses on predicting the delay in invoices based on the Machine Learning model, where the model studies the historical data and reduces the delays by proactive follow-up with the customers to ensure the invoice payments happen well within time.

A predictive model has been implemented to forecast if a consumer will remunerate the pending invoices well within the time, so that we can lessen interference measures on consumers, thereby lessening cost along with bettering the consumer relationships. The focal point was on individual invoices instead of customer level due to the non-availability of data.

Model performance can be increasingly improved by considering attributes like customer accounting period, seasonality, geographical data, along with accounting year.

Recommendations for further work: Further collection of customer-wise, region-wise data and detailed analysis is required. There is a high chance of improving the collection process for invoices thereby reducing the delays. This will also help in reducing the intervention actions between internal teams and customers thereby achieving better customer satisfaction.
Incorporating further details about the customer such as their income and gross profit can boost the learning dataset.

In the future, further to the probability obtained, investigation on algorithms basis grading can be done for amplifying business worth and utilize forecasting outcomes to enhance and accomplish collection fulfilment.

Bibliography

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

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

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*.

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

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

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

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

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

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

Investopedia. (2021).

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

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

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

Li, Ying., ACM Digital Library., Association for Computing Machinery. Special Interest Group on Knowledge Discovery & Data Mining., & Association for Computing Machinery. Special Interest Group on Management of Data. (2008). *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM.

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

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

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

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

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

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

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

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

Smirnov, J. (2016). *Modelling Late Invoice Payment Times Using Survival Analysis and Random Forests Techniques*.

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

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

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

W, B. A., A, K. K., & B, A. D. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences*, *115*(11), 2563–2570. https://doi.org/10.1073/pnas.1708279115

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

Zhang, Y., Wan, B., & Liu, W. (2016). *Recognition on Business Invoice_ZhangWanLiu*.

Zhu, J., He, H. J., Tang, J., & Yan, H. Z. (2016). Tax-control network invoice machine management platform based on socket. *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, 2343–2348. https://doi.org/10.1109/CompComm.2016.7925118

Appendix

Plagiarism Report

Prediction of delays in Invoice payments

by Aruna Kashinath

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

Prediction of delays in Invoice payments

ORIGINALITY REPORT

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

PRIMARY SOURCES

1	www.simonsfoundation.org Internet Source	1%
2	en.wikipedia.org Internet Source	1%
3	repository.tudelft.nl Internet Source	<1%
4	lib.buet.ac.bd:8080 Internet Source	<1%
5	Mohamed Hossam, Ahmed Ashraf Afify, Mohamed Rady, Michael Nabil, Kareem Moussa, Retaj Yousri, M. Saeed Darweesh. "A Comparative Study of Different Face Shape Classification Techniques", 2021 International Conference on Electronic Engineering (ICEEM), 2021 Publication	<1%
6	Submitted to University of Greenwich Student Paper	<1%
7	www.coursehero.com Internet Source	<1%

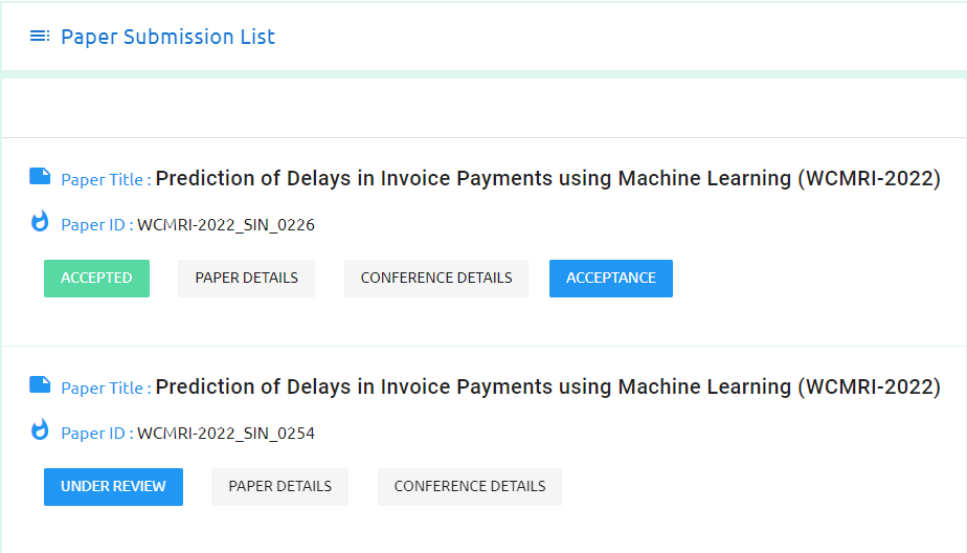
8	Submitted to Institute of Technology Blanchardstown Student Paper	<1 %

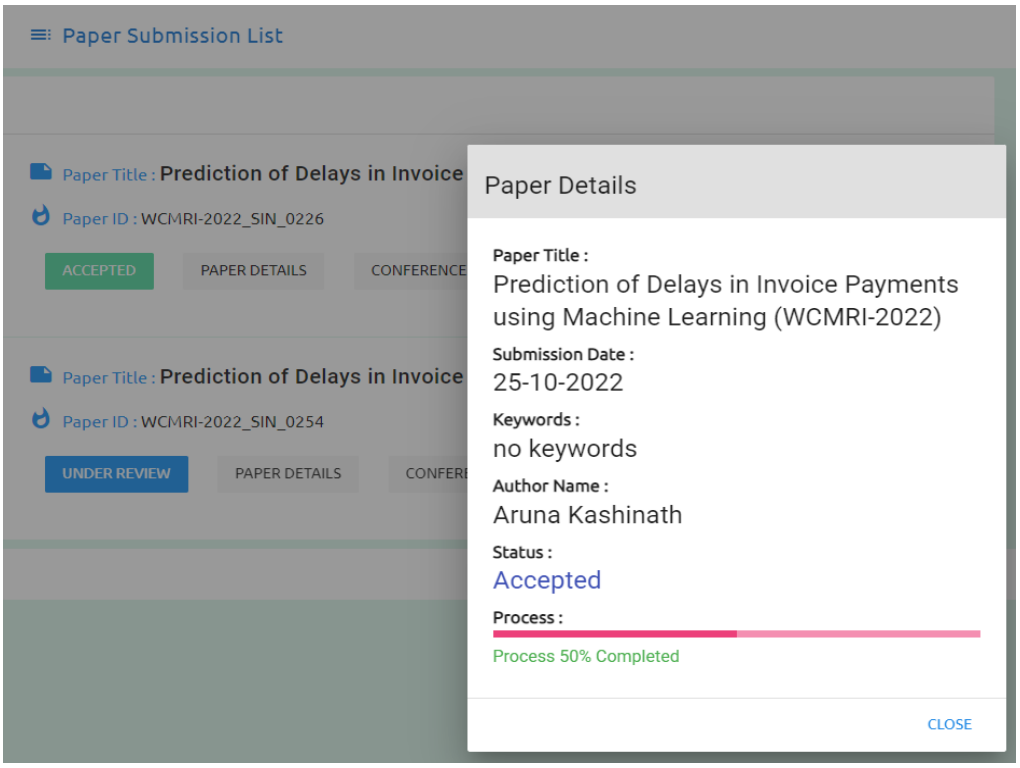
Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

Annexure
Publications in Conference

Paper submitted:

Aruna Kashinath, *Rashmi Agarwal,* "Prediction of Delays in Invoice Payments using Machine Learning" in **World Conference on Multidisciplinary Research & Innovation (WCMRI-2022)**. Submission Date: 25th October 2022.





Prediction of Delays in Invoice Payments using Machine Learning

Aruna Kashinath
REVA Academy for Corporate Excellence – RACE REVA University
Bangalore, Karnataka 560064
arunak.ba06@reva.edu.in

Rashmi Agarwal
REVA Academy for Corporate Excellence – RACE REVA University
Bangalore, Karnataka 560064
rashmi.agarwal@race.reva.edu.in

Mithun D J
REVA Academy for Corporate Excellence – RACE REVA University
Bangalore, Karnataka 560064
mithun.dj@reva.edu.in

Abstract*— Accounts Receivable (AR), is the most valuable asset of an organization. If it is not managed effectively, it can cause the firm serious financial hardships. 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. It is extremely important for large organizations that deal with thousands of vendors to fulfill their service level agreements with them to avoid penalties. The data is collected from a multinational organization consisting of the past two year's transactions. The research has been conducted by reviewing numerous papers related to invoice processing and different methodologies used to estimate consumer payment in AR. The purpose of this paper is to present a supervised modeling solution for predicting the payment outcomes of newly created invoices, thus enabling collection actions tailored for each invoice or customer. Since this is a classification problem, an ensemble method of Random Forest and Extreme Gradient Boosting algorithms has been applied and has achieved the highest accuracy if an invoice will be paid on time or not and can provide estimates of the magnitude of the delay which can improve the prioritization of customers and facilitates the daily work of collection personnel. It is estimated that the adoption of the model to prioritize the work of the Collection Management team will result in a substantial amount of savings. Consequently, the team would contact respective customers or account holders during the period in which the invoices are open, thereby reducing the organization's accounts receivable.

Keywords— Machine Learning, Invoice Processing, Invoice Payment, Delayed Invoices, Accounts Receivable, Predictive Modeling, Feature Engineering.

I. Introduction

Every business aims at ways to contain costs and increase the cash inflow. To achieve this, the organization's financial system must work with great efficiency. Within this financial system, AR classifies the costs to be paid by the customers or account holders.

As part of the booking process, an invoice is generated which must be paid according to the agreement with the customer. Prior to interference from the vendor or supplier organization, a 45-day limit has been sanctioned for open invoices. It is a common problem among many organizations that the customers fail to pay on time, and this results in subsequent follow-ups to remind their customers to pay the outstanding invoice amount [1]. As the number of invoices generated per day increases, it becomes increasingly important for the AR team to manually identify late payments or debts. To manage this process and to reduce the accounts receivables for the organization, the collection management team reaches out to its customers regarding their due payments to the organization.

Fig. 1 represents dealing with a large number of invoices from several customers in a month. Each bubble represents a set of invoices with the due date for that day, and the size of each bubble represents the number of invoices due on a particular day, from different clients. The position of the bubbles is the amount of money to be received on a given day. From Fig.1, it is difficult to prioritize which customers should be contacted first.

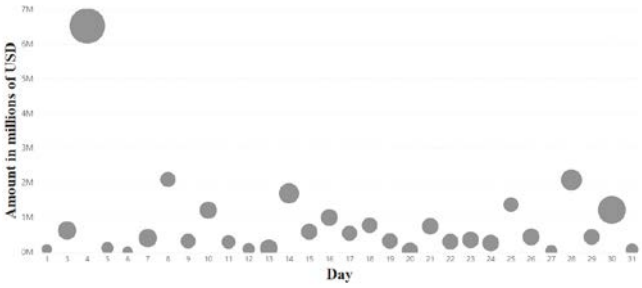


Fig. 1. Receivables over one month, distributed over payment data

The possibility exists of missing out on information such as which customers are more likely to be late in making their payments. It could be the easiest for the collection team to contact those customers that have a larger amount of money to be paid first since the organization wants to recover the invoices with huge payments. However, a larger sum of money does not necessarily guarantee higher-risk customers. Hence, predicting invoices that are most likely to be paid next can be the solution for prioritization of the right customers, better allocation of resources having a positive impact on cash flow estimation, thereby achieving financial stability. The focal point of the study in this paper is improving the efficacy of AR. Historical data of transactions have been utilized to identify patterns and create a predictive model. Using this model, new invoices will be categorized into payment buckets. Having this information can allow the collections department to concentrate on offending invoices along with bringing in proactive steps for speeding the invoice payment collection process.

II. STATE OF ART

There are few works of literature in the domain of invoice collection prediction. Some of the 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.

A. Late payment prediction

The study on late payment prediction could be done through machine learning as discussed in [2]. This was done by gathering invoices from four different firms including 2 Fortune 500 companies. Next to basic invoice information such as entry date, due date, and amount, the dataset also contained the delay of payment for every invoice ranging from 0 to more than 90 days. For this task, several decision tree algorithms were used such as PART and C4.5. The data was trained to be able to predict the size of the delay in terms of five classes: no-delay, 1-30 days, 31-60, 61-90, and more than 90 days. 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.

Following [2] and [3], using similar invoice and payment behavior features, the author [3] compared the results of five different models when trained on the dataset. These models are Decision Tree (DT), Random Forest (RF), AdaBoost (AB), Logistic Regression (LR), and Support Vector Machine (SVM). The conclusion from these papers suggests that 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 researcher 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. The author has also tested the models in two separate scenarios:

(a) Scenario One (Binary outcome): Predict whether an invoice is going to be paid on time (True/False).
(b) Scenario Two (Multiple outcomes): Predict whether the invoice belongs to one of four delay classes: no delay, short delay (within 30 days), medium delay (30-90 days), and long delay (more than 90 days).

The author [1] analyzed the characteristics of delayed invoices and problematic customers and concluded that there was no obvious correlation between invoice amount and delay. The author describes the use of machine learning for a finer-grained estimation task, namely spot factoring. The goal was to estimate the likelihood that an invoice will be paid in an acceptable timeframe. The paper describes three possible machine learning tasks for estimating the risk: binary classification for predetermined overdue days cut off; regression of the overdue days; and learning-to-rank which learns to optimize the risk-related ranking for the full range of instances. The outcome was a profit-driven evaluation that shows regression models can lead to higher profits and better spread the risk than classification and ranking models for spot factoring.

B. Customer Analysis

The author [1] did not predict if a particular invoice payment will be on time or late; instead, the focus was on the customer as a whole. Since a customer can have multiple outstanding invoices, the objective of the paper was to predict if the customer would likely pay on time or not pay at all. A new pureness measure was created to determine if a customer is good (pureness = 1) or bad (pureness = 0) to train their predictive models, by using features related to past on-time payment behavior and organization profile, rather than features related to past delayed payments. A model was built to predict those customers who have pureness between 0 and 1 (partially paid on time) and identify those who are likely to pay on time with a high probability, hoping to reduce the overall intervention actions taken. The past works were focused on increasing intervention actions on invoices that are likely to be late, while they focused on reducing intervention actions on customers who are likely to pay on time. Finally, using their pureness measure, they could determine the relationship between computed pureness and the predicted probability of paying on time.

C. Survival Analysis

Among the other methodologies, the study in the paper [5] uses Survival analysis as an approach as the thesis mentions that delayed invoice payment is not a classification problem because the interest lies when an event occurs rather than whether it occurs or not. Several articles deal with survival analysis [6-8]. The author mentions the predictive models are fit using large historical sets of existing customer data that extend over many years; default trends, anomalies, and other temporal phenomena that result from dynamic economic conditions are not brought to light. They have introduced a modification of the proportional hazards survival model that includes a time-dependency mechanism for capturing temporal phenomena and developed a maximum likelihood algorithm for fitting the model. Using a very large, real data set, they were able to demonstrate that incorporating the time dependency can provide more accurate risk scoring, as well as important insight into dynamic market effects that can inform and enhance related decision-making.

The author [5] views the invoice late payments as regression rather than a classification problem and survival analysis fitted their purpose. Survival analysis and a novel ensemble method of Random Survival Forests were applied to the right-censored data of late invoices. They proposed two separate models, for first-time debtors and for repeated debtors, and explored the effect of different predictors in a model. Random Survival Forests proved to have advantages over the Cox Proportional Hazards model as there were no underlying assumptions that needed to be taken into consideration. Overall, it is concluded that the Random Survival Forests model which additionally uses the historical payment behavior of debtors, performs the best in ranking payment times of late invoices.

D. Prediction based on Account Payables (AP)

On the other end of the spectrum, there is a study by [9], 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.

Considering the learnings from various papers, the study in this paper 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.

III. PROBLEM DEFINITION

Collections Management is responsible for collecting payments against open invoices within the organization considered for the study. Some of the critical concepts in collection management are reviewed: in a transaction contract, the payment term is defined as days provided for an invoice transaction cost being paid (from invoice created date to due date), usually, 30, 45, or 60 days and the due date is the maximum time when the transaction cost must be paid. An overdue (or late) invoice is a bill that is late in its payment, i.e., the invoice's due date is passed, and the customer has not yet made the payment whereas, in contrast, outstanding invoices are not paid yet but are not late either, their due dates lie in future. Invoice can also be paid in advance, earlier than the agreed due date. The customer is typically at risk when making these advance payments as they can pay for the service or good, but not receive it at the end [10].

The prediction of invoices is a typical classification problem using supervised learning, where, in the given historical dataset, the resultant set should include invoices' features, building a machine learning model to perform the classification of invoices as on-time or late with a regression model to predict the estimated delay.

IV. DATA SOURCE

The analysis in the following sections is based on a fortune 500 company's invoice data set that provides technology services. The dataset contained 34,752 invoice data, which processes around 1500 invoice transaction data in a month. For confidentiality, customer-wise details were not available for study and hence only invoice level details were considered. Most fields in the dataset were categorical features that posed challenges in terms of variables with no numerical relationship between levels. The most important numerical feature is the value of the transaction in USD. The cost of an invoice transaction was between <1 dollar to ~20M$. There are indicative dates such as the Transaction Date when the invoice was created and the Due Date, calculated as per the terms of invoice payment due.

The invoices have been categorized according to the number of days that they have been delayed as showcased in Fig. 2. It is evident from the graph that over 50% of the delays are contributed from the buckets 16-30 and 31-45 days together. There is a need to pay attention to these invoices and try to find common features shared by these invoices.

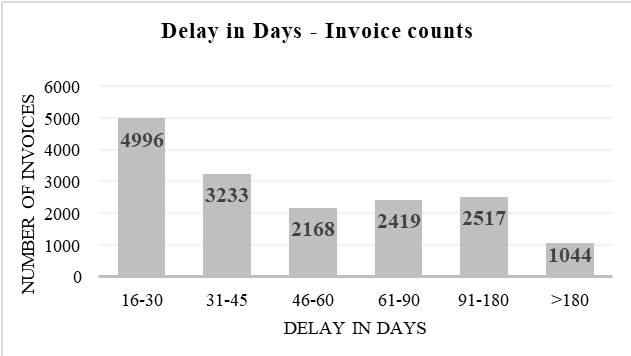


Fig. 2. Invoice Payment Delay in Days

Out of the dataset considered, 53% of invoices had a positive outcome, meaning these invoices were closed without delays and 47% had delays in payment.

The invoices paid before the due date are divided into 2 classes as shown in Fig. 3. The invoices can be classified into (-1) for early payments, and (0) for on-time payments.

Out of 53% of invoices that had no delays in payment, over 47% of invoices were paid earlier than the due date and 6% were paid on time. It is depicted in Fig. 3 showcasing the distributions of the 2 constructed targets, by the number of invoices.

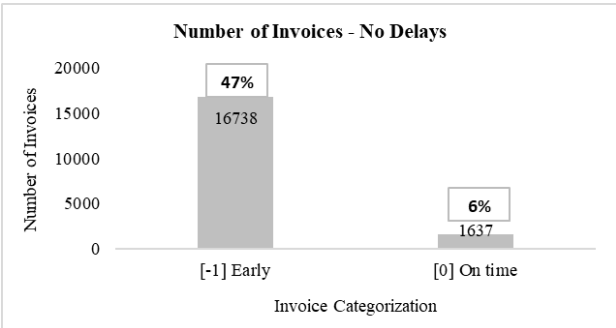


Fig. 3. Invoice Categorization – No Delays

Based on account receivables management best practices, the delays are in turn classified into late (1) if the payment fall between 1 to 30 days after the due date; very late (2) if the payment fall late between 31 and 90 days after the due date; and critically late (3) if the payment falls later than 91 days and onwards as showcased in Fig. 4. It is important for the vendor to identify those customers likely to pay earlier than the due date on the contract to define collection strategies.

Out of 47% of invoices that were delayed for payments, 29% were up to 1 month late, 11% were between 1 to 3 months late, and 7% were delayed for more than 3 months. Depicted in Fig. 4 are the distributions of the three constructed targets, by the number of invoices. Invoices that are delayed for much longer are the most problematic and therefore called "Critically late" invoices.

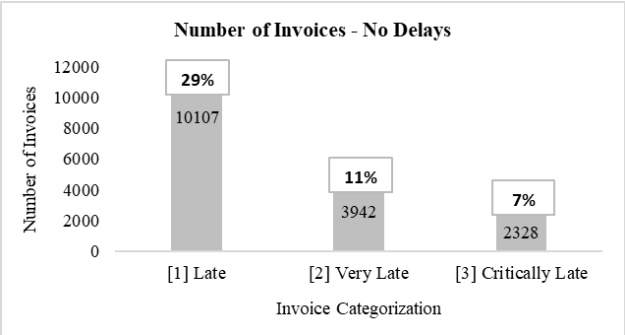


Fig. 4. Invoice Categorization - Delays

The dataset contains information related to invoicing payments and other categorical variables pertaining to e-invoicing, web invoicing, Purchase Order (PO) mandate, etc. This dataset presents a significant challenge due to the absence of customer-relevant information. Due to privacy constraints, the scope of the project was reduced only to invoice details.

Feature Engineering was performed to derive relevant features in order to build a machine learning model. Out of 23 features, only 14 have been selected as candidates for the feature engineering part. In short, irrelevant columns that do not represent valuable information to the problem were dropped. Previous literature reviews on AR helped derive the most valuable features from the data. Categorical columns with too many levels are also irrelevant. In fact, their use will undermine the model's performance as there will be fewer examples available for each level of training, which will reduce the ability of each level to have any impact on the model's outcome [10]. There was no missing data in any of the columns and hence there was no need for further imputation to the data. Invoices for which payment is delayed or advanced over the course of one year are rare and usually involve special contracts that should be considered separately. Label encoding and one-hot encoding were done to the categorical variables to create new attributes on basis of distinctive features in the attribute set.

V. PROPOSED METHOD

This paper aims to provide algorithmic solutions to the team in predicting the delay in invoice payments based on the features like reasons for which the invoice is created, Line of Business (LOB), invoice amount, type of invoices whether e-invoicing or web, Purchase Order Mandate, and the Country in which the customer is located. Based on proactive predictions, delays in invoice payments can be reduced.

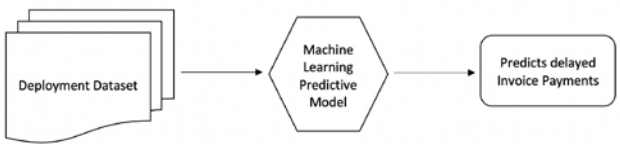


Fig. 5. Algorithmic Solution for Delayed Invoice Payment

The solution for the delayed invoice payments has been depicted in Fig. 5. Dataset is sourced to a supervised machine learning model which predicts delayed invoice payments along with the probability of delays. In this dataset, we have close to 50% defaulting on the payment. With historical examples of invoices and their end result, the paper aims to answer the following questions for a new invoice:

(i) Will the invoices be paid on time or late?
(ii) What could be the duration in which the invoice will be paid?

An attempt is being made to develop a classification system to identify the category a new invoice belongs to, based on a training set of data that includes cases with known outcomes. As a supervised learning problem, formulate the invoice outcome prediction problem given the consequence of the historical invoices, a model is created for forecasting the payment time for a new invoice.

VI. MODELING

As outlined before, this is a binary classification problem to predict whether an invoice will have a payment on time or delayed. Data discussed in the previous section has been fed into multiple models to classify if invoice payment is done on time or delayed. Based on the problem statement and the data availability, a supervised learning approach to train the classifiers are listed in Table I wherein each of these classifiers are evaluated and used to improve our results as different models will be better suited for the dataset considered for the study of this paper.

TABLE I. LIST OF VARIOUS CLASSIFIERS EVALUATED IN THIS PAPER

Classifier	Approach
XGBoost (Extreme Gradient Boosting) Classification	An ensemble of decision trees using gradient boosting classification.
XGBoost Regression	An ensemble of decision trees using gradient boosting regression.
Random Forest Classification	Ensemble of decision trees combining ideas of a random selection of features and bagging.
Random Forest Regression	Ensemble of decision trees combining ideas of a random selection of features and bagging.
Neural Network Classifier	Trained a neural network with 3-8 layers with binary cross-entropy loss.
K Neighbors Classifier	Uses proximity to make classifications or predictions about the grouping of an individual data point.
Linear SVC (Support Vector Classifier)	Fit to the data returning a "best fit" hyperplane that divides or categorizes data.
Multi-layer Perceptron (MLP) Regressor	A feedforward neural network training system that implements a multi-layer perceptron regression algorithm.

VII. MODELING EVALUATION

The objective of this section is to identify metrics and evaluate different machine learning models for the prediction of invoice late payments.

A. Training

The dataset considered for the study is split into two categories that are 75% of training and 25% of the test set. The training dataset consists of a "Paid" label with the status "Yes" or "No" indicating if there is a delay in payment. The testing set is used to evaluate the performance of classifiers. Regression is considered to predict the magnitude of delay.

B. Metrics

Evaluation metrics such as Accuracy, Precision, Recall, F1 Score, Receiver Operating Characteristic curve (ROC) for classification, and Mean Absolute Percentage Error (MAPE) for regression have been derived from the models. The aim is to achieve high precision and high recall on invoices paid late as there was no action needed for invoices paid on time. High precision would indicate that most of the invoices labeled as "paid late" are indeed "paid late". High recall would imply that the model is able to detect the majority of invoices that are going to be "paid late". A confusion matrix was used in the study to compare the number of predicted instances in each class to the correct classification.

C. Predicted Features

The predicted features in the study are the "Paid" label that classifies if an invoice transaction cost was paid late or not and the "Days Paid Late" label predicts the magnitude of delays.

VIII. RESULTS

In this section, the result of the modeling methods considered to predict the late paid invoices is discussed. In summary, "Boosted Trees" and "Random Forest" performed the best based on the Accuracy and F1 Score in Table II. The classification models were considered to predict the payment as on-time on late.

In order to predict the magnitude of delays in payments, XGBoost Regressor, Random Forest Regressor, and MLP Regressor were considered. Mean Absolute Percentage Error (MAPE) and Coefficient of Determination were the metrics considered for regression models. MLP Regressor performed the best with a MAPE of 3%.

TABLE II. MODEL EVALUATION

Models	Accuracy	F1 Score
Logistic Regression	68.62%	58.02%
XGBoost Classification	81.57%	69.50%
Random Forest Classification	80.60%	70.12%
Neural Network	36.73%	22.42%
K-Neighbors	45.02%	28.22%
Linear Support Vector Classifier (SVC)	55.96%	32.64%

The Confusion metrics of the sample test are shown in Fig. 6. The model is able to classify True Positive with values as 1005 indicating that these invoices will be paid on time and has been able to classify 675 invoices as False Negative indicating delays, however, are paid on time.

Confusion Matrix for test result		**Prediction**	
		On Time	Delay
Actual	On Time	1005	675
	Delay	530	1266

Fig. 6. Confusion Matrix for test results

The Precision Score is 0.69 for the model indicating that when it classifies the invoices as late, it is correct 65% of the time. The recall is 0.70 indicating that 70% of the time the model is able to classify the late payments as delayed. F1 Score being 0.67 states the equilibrium between precision and recall.

The Invoice Cost, Line of Business (LOB), and Credit Hold were considered Key Drivers/Feature Importance causing the delays. Credit Hold is a consequence for a customer who is consistently late in making payments, has exceeded their credit limit, or is identified as a bad risk.

The accuracy of a classifier to distinguish two classes of data is a measure of the ROC curve. Fig. 7 shows the ROC curve for this study with an accuracy of 73%.

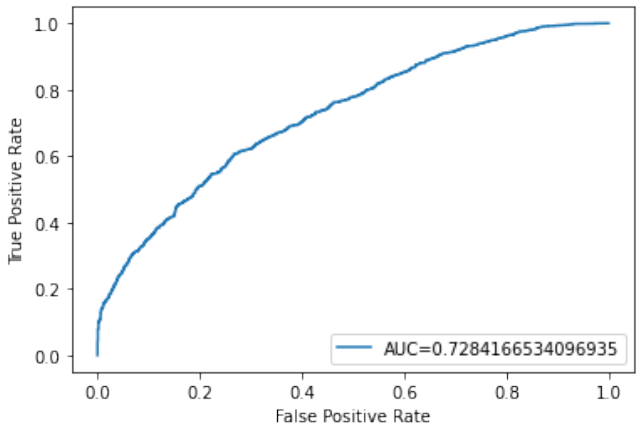


Fig. 7. ROC Curve

IX. CONCLUSION

The paper focused on the classic binary classification task and the regression for the number of days overdue. Both XGBoost and Random Forest Classifier stood out as outspokenly good performers. MLP Regressor estimated the magnitude of delays with a very minimal error rate described as the best regressor model. The model prediction should be used to focus on defaulters instead of reaching out to a customer who pays on time. The organization can save up to 40% cost by not contacting those customers who will pay on time.

X. FUTURE WORKS

The objective of this paper is solely to predict delays in invoice payments based on Machine Learning models, which analyze historical data and eliminate delays by proactive follow-up with customers to ensure payment is received on time.

A predictive model has been implemented to forecast if a consumer will remunerate the pending invoices well within the time, so that we can lessen interference measures on consumers, thereby lessening cost along with bettering the consumer relationships. The focal point was on individual invoices instead of customer level due to the non-availability of data. Model performance can be increasingly improved by considering attributes like customer accounting period, seasonality, geographical data, along with accounting year.

A further collection of customer-wise, region-wise data and detailed analysis is required. The collection process for invoices can be improved, thereby reducing delays. This will also help in reducing the intervention actions between internal teams and customers thereby achieving better customer satisfaction. Incorporating further details about the customer such as their income and gross profit can boost the learning dataset.

XI. REFERENCES

[1] R. Stahlbock, G. M. Weiss, and M. Abou-Nasr, *Proceedings of the 2018 International Conference on Data Science: ICDATA '18*. 2018.

[2] S. Zeng, P. Melville, C. A. Lang, I. Boier-Martin, and C. 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*, 2008, pp. 1043–1050. doi: 10.1145/1401890.1402014.

[3] H. Peiguang, "Predicting and Improving Invoice-to-Cash Collection Through Machine Learning," 2015.

[4] L. Coenen, W. Verbeke, and T. Guns, "Machine learning methods for the short-term probability of default: A comparison of classification, regression and ranking methods," *Journal of the Operational Research Society*, vol. 73, no. 1, pp. 191–206, 2022, doi: 10.1080/01605682.2020.1865847.

[5] J. Smirnov, "Modelling Late Invoice Payment Times Using Survival Analysis and Random Forests Techniques," 2016.

[6] M. Stepanova and L. Thomas, "Survival Analysis Methods for Personal Loan Data," *Oper Res*, vol. 50, no. 2, pp. 277–289, 2002, [Online]. Available: http://www.jstor.org/stable/3088495

[7] J. J. Jaber, N. Ismail, U. Kebangsaan Malaysia, and M. Siti Norafidah Mohd Ramli, "Journal of Internet Banking and Commerce CREDIT RISK ASSESSMENT USING SURVIVAL ANALYSIS FOR PROGRESSIVE RIGHT-CENSORED DATA: A CASE STUDY IN JORDAN," 2017. [Online]. Available: http://www.icommercecentral.com

[8] J.-K. Im, D. W. Apley, C. Qi, and X. Shan, "A time-dependent proportional hazards survival model for credit risk analysis," *Journal of the Operational Research Society*, vol. 63, no. 3, pp. 306–321, Mar. 2012, doi: 10.1057/jors.2011.34.

[9] T. Tater, S. Dechu, S. Mani, and C. Maurya, "Prediction of invoice payment status in account payable business process," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11236 LNCS, pp. 165–180. doi: 10.1007/978-3-030-03596-9_11.

[10] H. C. Pfohl and M. Gomm, "Supply Chain Finance: Optimizing Financial Flows in Supply Chains," 2009.