Tentative Payment Date Predictor

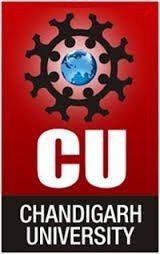
Submitted in partial fulfillment of the requirements for the award of degree of

## DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

**Chandigarh University, Gharuan**

### BACHELOR OF ENGINEERING IN

**COMPUTER SCIENCE & ENGINEERING**



#### Submitted By:

**SHIVAM ( 18bcs3054 )**

**Mentor Signature****:**

1. **Abstract**

Keeping a steady cash flow is one of the biggest if not the biggest problem that Small to Medium Enterprises (SMEs) deal with daily. Within the different types of cash flow, Accounts Receivable (AR) classifies the balance of money that needs to be paid by the company’s customers. In the most typical case, after receiving goods or services, the customer receives an invoice with the amount that is owed to the supplier. However, this often does not happen before the aforementioned date, meaning that the invoice is often paid late. Intervention requires resources and over-intervention could cause unwanted customer dissatisfaction. Knowing whether an invoice is going to be paid late can be vital information. Current methods of late payment prediction focus only on the history between the seller and the buyer and are unusable when this history is not present. Intuitively, one’s business depends on the relationships and transactions that it has with its neighbors. Suggesting that neighbor behavior could be useful when predicting the cash flow of a company. Unfortunately, this type of information is not always given and needs to be data mining from non-relational data. This work presents

a method for building a relational network of SMEs using entity resolution and improving the current state of the art of late payment prediction using features extracted from the graph

# Introduction

The tentative payment date predictor is a B2B project which tends to predict payment dates from companies using previous data sets. This project uses machine learning models to construct logic on the data that is available from the previous records and construct a model which uses data for prediction. A B2B model is a type of business where a company doesn't deal with the real customers directly but it provides services to the companies and exchange business with them.

This type of business run on a credit system which means that goods are given to the companies and shops on a postpaid basis and the amount of time a shop or company takes to pay the money in a given span of time affects their credit score in the market and their cap of product withdraw from the market. We have made this project for the companies giving products on credit that so that they can have safe play with companies that would tend to pay their loan on time and the companies who won't and this would help company in doing fair business with minimal loss.

## Currently available solutions

Companies such as Oracle and SAP have created software that can process invoices and automatically take actions as a result of that. In the case of Microsoft, users can install a plugin that gives them more insight into their sales invoices . The tool provides a prediction on whether a specific invoice will be paid on time. The tool categorizes the invoice into two prediction classes; on-time and delayed. Attached to this prediction a confidence level is given. The confidence levels go from Low, Medium to High. The levels correspond to the 70%, 80%, and the 90% confidence thresholds respectively. Being a generic tool, the model has been trained on a range of small and medium businesses. Made to be able to serve different types of companies when it comes off the shelve. The model improves over time by using the user’s data to retrain. Resulting in a model that will eventually be fitted on the data of the user. While the complete

architecture of the solution is unclear, it does not seem to account for the possible transactions between the parties in the system. Rather, the predictions are made by solely looking at an individual customer and its history. Intuitively, one’s business depends on the relationships and transactions that it has with its neighbors and that growth or bankruptcy does not happen in isolation. In some cases the customer becomes insolvent for a brief period, meaning it is not able to pay its debts. This will impact the outstanding sales, despite the customer being trustworthy or not. UK’s association of business recovery professionals explains that around 27% of insolvencies are triggered by the insolvency of another company. Adding that there is some type of "domino effect" in play. This suggests that a company’s ability to pay off invoices is dependent on the insolvency of its suppliers and customers. Which in turn, also depends on further relationships. While this relational information can be beneficial, this type of data is not always available. Moreover, in cases where it is, the data is not guaranteed to be in a standardized form, rich enough that it can be used to construct a reliable network.

Un standardized dataset, entity resolution can be used to find the unique companies or entities in the dataset. These data mined entities can, in turn, be used to create the nodes of the desired networks. With edges in the network representing the transactions between these companies.

Once the network is constructed different methods can be used to extract features from the graph. Besides features such as average degree and edge weights, in the last several years it has become more and more popular to use embedded representations of the network as features. The reason this needs to be done is that graphs cannot directly be used in machine learning algorithms, as this data needs to be Euclidean to do so. While the graph embedding concept is not new, the recent breakthroughs in deep learning and more specifically computer vision and representation learning, have inspired new methods that make graph embeddings very efficient and scalable.

Making these methods usable on networks with billions of nodes and edges. While showing promising results fields of recommendation and drug discovery. Exact To make this research possible, the experiments are performed using the data provided by Exact. Exact is a software company that is specialized in software for accounting, ERP, CRM and other types of software for Small to Medium Enterprises . As of writing this report, the online platform has 400k users with the majority being in the Netherlands. It is estimated that approximately 20% of Dutch SMEs use Exact software. Exact strives to improve its products by incorporating features that help its customers grow. Because of the large impact that cash-flow has on these businesses, it is in Exact’s best interest to help their customers manage their receivables. Chapter 4 will describe the software and how it is used in more detail.

# Feasibility Study:

The described problem can essentially be split into two major parts: 1. Creating a network of Dutch SMEs from unstandardized and noisy data. 2. Improving the current methods of late payment prediction using features extracted from the network. The thesis also provides an end- to-end solution, of how data about the business supply chains can be used to build a network of SMEs through entity resolution. Furthermore, it shows how this network can be leveraged through methods such as graph embedding, to improve the predictions of late-payments. The focus of this thesis is to see whether the addition of features extracted from a graph of related companies can improve the accuracy of late payment predictions. To do this we define the following research question and underlying sub-questions.

* MQ: Can graph features be used to improve the prediction of late invoice payments compared to currently popular methods? This is done by answering three separate sub-questions:
* SQ1: How can a network of SMEs be built from data that is unstandardized, noisy and partial?
* SQ2: How should the data and the graph be structured to be able to extract meaningful features?
* SQ3: Does the addition of graph features improve the prediction of late payments?

The defined method and experiments all try to answer one of the sub-questions. An overview of the experiments and which sub-question they try to answer can be found in chapter

# Methodology/ Planning of work

## Late payment prediction

Late-payment prediction models are used on a more granular level, predicting when specific payments (i.e invoices) are going to default. While credit scoring is a well-researched topic, there has been very little research done in regards to predicting late payments

## Financial transactions

Financial transactions can be modeled as a dynamic graph to analyze the interaction between different financial bodies. Work done by [33] explores different types of metrics in an economic system model as a complex network. The network explains monetary transactions between 105 clusters, each representing an economic activity standardized by the UN. The paper provides the following two contributions: A Network definition that is as follows:

* Node, is an economic activity cluster, with the node weight being the summed transactions within the cluster.
* An undirected Edge is present when money flows between two sectors. Its weights show the summed money flow between two clusters in either direction.

## Graph analysis and feature engineering

Graph embedding methods have seen a spike in interest and application in the last couple of years. These graph embeddings make it possible to encode graphs, making it possible to use graphs as input in various machine learning algorithms.

Generally, the embedding algorithms are categorized by its method:

* Matrix Factorization: the embedding is achieved by factorization of the adjacency matrix.
* Random Walk based Deep Learning: uses the Skip Gram architecture to learn effective embeddings of random walks generated from the graphs.
* Non-Random walk Deep Learning: these methods leverage network architectures such as autoencoders or graph convolution layers to embed the input

#### Feature engineering

Since networks cannot directly be used as input in machine learning models. The problem of prediction relies primarily on the quality of engineered features. Therefore, it is important to have effective techniques that extract meaningful features from the networks. A well-known problem in this domain is the problem of link prediction, where the goal is to find missing links in a network using information about the nodes. Mutluetal.

# Entity Resolution

Entity Resolution (ER for short, also known as Entity Matching, Entity Disambiguation, Record Linkage) describes the problem of finding unique entities from either single or multiple data sources. The paper done by Konda et al. describes that while there has been an effort made in understanding the problem there is very little to no published work on ER in practice, end-to- end. The general outline of the paper is to show the methodology and workflow of doing ER in a real-world scenario. It argues that every unique case needs experts to differentiate the records and heuristics. The paper contributes a description of a real-world application, the goals set by the stakeholders involved and a description of the common ER challenges in real-world applications. The authors describe the first step to be setting up the matching rules:

1. Two records are a direct match if the unique ID is the same in both records.
2. If Titles are similar. 3. If similar individuals are involved.

# Module & Team Member wise Distribution of work

Ravi Kumar ( 18bcs2695 ) : Project planning, Data collection and Model testing Ritik Kumar ( 18bcs6659 ) : EDA, Data Framing and Project co-lead

Sumit Kumar ( 18bcs6654 ) : Feature Engineering, Model implementing and Project lead

# Software and Hardware Requirements

1. Python
2. Pandas
3. Numpy
4. Jupyter notebook