UNIVERSITY OF TARTU

INSTITUTE OF COMPUTER SCIENCE

Software Engineering Curriculum

Madhu Tipirishetty

Predictive process monitoring for Lead-to-Contract process optimization

Master's Thesis

Supervisor: Peep Küngas, PhD

Tartu 2016

Predictive process monitoring for Lead-to-Contract process optimization

Abstract: Business processes today are supported by enterprise systems such as Enterprise Resource Planning systems. These systems store large amounts of process execution log data that can be used to improve business processes across the organization. The process mining methods have been developed to analyze such logs, which are capable of extracting process models. These methods, in turn, have been applied in conjunctions with predictive monitoring methods for early differentiation of desired and undesired outcomes. Although predictive monitoring approach has recently caught attention and found application in recommendation engines, which suggest cases to improve business process outcomes, there is no much research on how contextual data, such as clients financial indicators and other external data, may improve the quality of recommendations. This thesis examines whether including the external data with the event data affects the accuracy of predictive monitoring for early predictions positively. More specifically, this thesis reveals usage of context data had the adverse effect on the performance of learned models. Furthermore, the study indicated that the usage of first three events from the event logs with internal data is sufficient to predict the label of an opportunity in the sales funnel.

Keywords: Predictive monitoring, Machine learning, Context-aware data

CERCS-code: P170

Müügiprotsessi optimeerimine läbi ennustava protsessiseire

Lühikokkuvõte: Äriprotsesside toetamiseks on üha laiemalt kasutusele võetud ettevõtte ressursside planeerimise (ERP) tööriistad, sealhulgas CRM süsteemid müügiprotsessi jaoks. ERP süsteemid salvestavad oma töö käigus protsesside logisid, mille oskuslik käsitlemine võimaldab efektiivistada äriprotsesse. Protsessilogide analüüsimiseks on välja töötatud protsessikaeve meetodid, mis oskavad logidest pöördprojekteerida tegelikult käivitatud protsesside mudelied. Neid meetodeid on rakendatud koos ennustava seire meetoditega protesside tulemuste soovitud ja soovimatute tulemuste varajaseks tuvastamiseks.

Kuigi ennustav seire on hiljuti rohkelt tähelepanu saanud ja leidnud rakendamist soovitusmootorites, mis pakuvad välja soovitusi äriprotsesside parendamiseks, ei ole seni palju uuritud kontekstiandmete, nt müügisüsteemi kirjetes klientide finantsandmed, mõjust ennustava seire tulemustele soovituste kontekstis.

Käesolevas magistritöös uuritakse kontekstiandmete mõju ennustava seire mudelite kvaliteedile müügiprotsessi optimeerimise kontekstis. Eksperimendid näitavad, et välistel kontekstiandmetel on pigem negatiivne mõju, samas kui sisemistel, protsessi käigus kogutud kontekstiandmetel on positiivne mõju mudelite kvaliteedile. Muuhulgas lähtub eksperimentidest, et juba kolme esimese sündmuse baasi saab müügiprotsessis ennustada müügi õnnestumist.

Märksõnad: Ennustav seire, masinõpe, kontekst

CERCS-code: P170

Contents

1 Introduction

The finished master thesis presented here is the part of Software Engineering master at the University of Tartu carried out in Faculty of Science and Technology.

1.1 Thesis context

Business processes play a pivotal role for the companies emerging in this competitive world. Enterprise Resource Planning systems such as Customer Relationship Management(CRM) systems are gaining a huge impact on the market. These business processes have enormous amounts of data via the process and execution logs adopted for the improvement of a process. Such process records have enough data which are employed with process mining methods for extracting the models. In addition to this, advanced techniques like predictive monitoring can be applied in combination with the process mining to yield better results. Although these advanced techniques are available, there is no substantial research in the recommendation solutions for the financial indicators.

Sales pipeline in lead-to-contract(CRM) processes are ambiguous and time-consuming in this fast paced environment which in turn leads to the poor process quality, and there are more chances of losing the potential sales because of the faults in the processes. Errors in the process can be for many reasons such as spending extended duration on cases, a sales representative in the company might do things in an unusual way, etc. All the information related to the user activities, events, durations, etc. captured in process execution logs can be used to bring improvement in the sales opportunities.

If the recommendation of cases is provided on hand for the incomplete cases to the sales funnel, then there are higher chances of attaining the opportunities. Although there is a tremendous increase in the growth of the predictive analytics related to enterprise software, there is less research related to predictive monitoring of context aware process data for early predictions. In this thesis work, we studied the applicability of external data with the combination of event logs for predicting the labels of the opportunities in the sales funnel.

Our study indicates that the usage of the external data indicators had the unfavorable effect of the performance of the learned models. However, internal data collected throughout the sales process gave the best performance. Finally, the study indicated that the three first events from the event logs are sufficient to predict the label of the opportunity in the sales funnel.

For ranking the opportunities, in the sales pipeline, we used the event logs with the first three events combined to internal data for predicting the current running cases based on the probabilities of positive and negative ratios of the opportunities obtained from the classification model.

1.2 Thesis problem and scope

The primary research goal of this thesis is to deliver a ranking system for sales funnel based on the lead to contract process. The existing process analytics as a part of predictive analytics do not completely solve this problem of generating recommendations based on logs. The proposed recommendation solution should be able to identify the deviance in processes and suggest the cases which have a high probability of close ratio to optimize the sales and business process.

The main research question is to design and implement a ranking system of cases, which will increase the performance of the lead-to-contract process. More specific research question are,

- Which predictive monitoring methods should be applied to distinguish and predict the desired and undesired outcome of lead-to-contract process execution by inspecting application event logs and external context data about workflow objects?
- How to use predictive monitoring methods for early prediction of cases?
- Research objectives are to identify existing methods, techniques, and tools, which will support the continuous construction of ranking the cases as early as possible.
- Develop methods for extracting interpretable rules and patterns that distinguish positive vs. negative outcomes in a business process. Validate the methods on lead-to-contract.

1.3 Outline: Predictive Monitoring Approach .

Predictive process monitoring based on the past process executions which analyze event logs as the input data for training the completed cases and evaluates the unfinished cases with the given predicate.

Based on the literature review of the state of art techniques explained in Section 2 [1], we considered the following criteria of predictive monitoring approach represented in Figure 1. Event logs are data encoded into various formats to check for the optimal features described in Section 4. After the successful preparation of target data from data preprocessing stage, finished cases are trained with the "Random Forest" method. Based on the training results obtained from the above step, unfinished cases are classified, and labels are predicted based on the probabilities acquired from the predictions. Then the ranking of cases is provided to the sales.

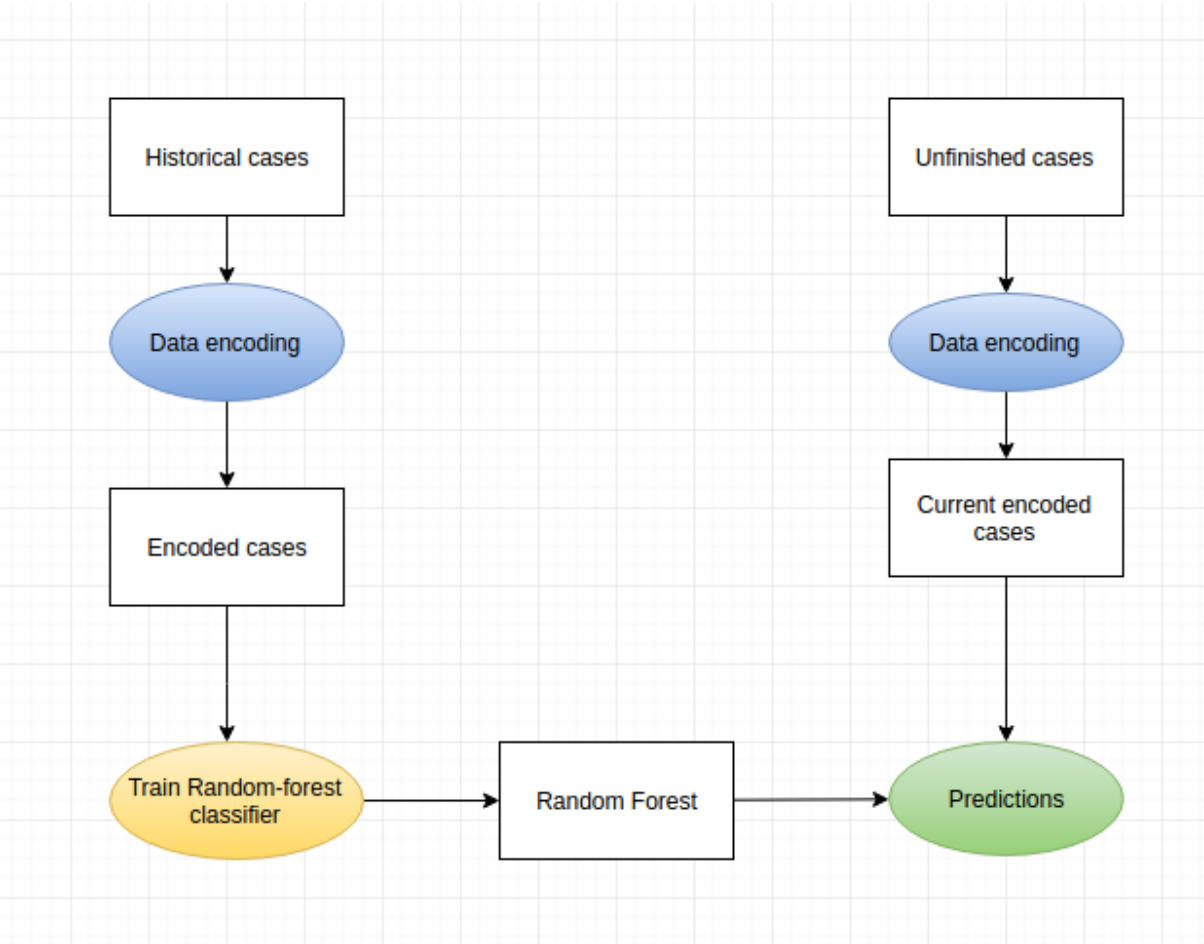

Figure 1: Predictive monitoring approach [1]

In Figure 1, historical cases are the closed cases from November-2015 to April-2016. In data encoding, data is transformed into different formats to derive the optimal feature set which is described briefly in Section 4.3. After the encoded data processed, they are unified with distinct event folds. "Random Forest" method is used in our study based on the experiments defined in Section 5.3. Unfinished cases on the other side are the current running cases which encoded in the same way the completed cases were encoded to get the predictions.

In the next chapters of this thesis, we precisely describe the related work which gives the overview of the state of art techniques in Section 2. Then in Section 3, we provide an insight of the background to understand the lead-to-contract process, the concept of process mining and its usability in our study, and the description of tools used in this thesis. Section 4 contains the data preprocessing with a brief introduction of datasets used in our study. In Section 5 we explain the experimental settings used in our work such as selection of data features, sequence lengths, and model.

2 Related work

Business process management systems have got the attention in the modern business perspective with continuous improvement of the process via process mining applications. Context-aware monitoring of business process in Enterprise Resource Planning(ERP) systems such as CRM using process mining techniques corresponding to predictive monitoring of early case predictions is a new research area. The following section showcases the review of the essential process mining techniques like predictive monitoring for early predictions, recommendation generation via sequence classification, the concept of predictive monitoring and deviance mining utilized in the current market scenario.

"Leontjeva, A. used complex sequence encodings via predictive monitoring for business processes" [1]. Firstly historical traces were encoded in complex sequences. Next, the Random Forest training was applied for the closed cases for predicting the outcomes of the cases. Here the completed cases are considered for the classification with a label differentiating whether the case is positive or negative. For the unfinished sequences, the data was encoded in the same way using complex encodings. After training the cases, half-done cases are predicted with the outcome. In this research, the data encodings were done in four ways, using boolean encoding, frequency-based encoding, simple index encoding and index-latest payload encoding. For evaluating this model, patients and insurance data was studied as the input. We used this work to carry our experiments and constructing our datasets.

Di Francescomarino, C. [2] classified the approach in two steps starting with clustering and classification to make predictions. Historical traces encoded in the form of frequency and sequence matrices; then they were clustered with similar groups. Similar group clusters selected for classification. Classification algorithms like random forests and decision trees applied in the next phase. For predictions of unfinished cases, clusters groups are selected to which the trace belonged to and based on the decision trees; predictions generated. Finished cases trained, and incomplete cases are classified to predict the case predicate. In this research for the experimental purpose, Weka was used for clustering and classification of decision trees using Weka J48 implementation of the C4.5 algorithm. This work differs with respect to the clustering in the present thesis work as it doesn't fit our data.

Verenich, I. [3] followed the similar approach mentioned earlier with few changes. Completed cases were encoded using the complex symbolic sequences. In the first phase, running traces were clustered and in the second step, for each group, cases are trained using random forests method. To predict the outcome of the unfinished sequences, for the running track, closest cluster based on Euclidean distance is selected, and respective classifier was applied. For Clustering Hierarchical agglomerative clustering and k-medoids clustering were used in this research. This paper differs our selected approach by clustering which was not included in our study.

"Van der Aalst, W. M. used time predictions based on process mining"[4], completion times of a process instance is predicted for the incomplete cases based on the closed cases. For the event logs, transition system is generated using the sequences and they are annotated for the predictions. Transition system has elapsed times, total execution times

and average execution times of the cases. The partial traces mapped to the transition system. The implementation provided via ProM plug-in. This work reminds us about the durations of the cases in the event log analysis.

Sebu, M. L. [5] developed a platform to diagnose the event logs and identify the different traces to send the alerts to the process manager to improve the process performance. Event logs in XML, XES format data were considered as the input for this model. Their model has four services, namely configuration service, data collection service, event processing service and delivery service. Configuration service has the rules in the description and the alerts. If the rules were not matching with the trace, then the alert is sent. Data collector service exports all the events into the XML and added to the queue for rule verification. Event processing service will select events from the queue and prepares the traces from the events and check whether they are matching with the traces defined with the rules. Delivery service used for providing the alerts whenever there is an unnatural state. This work reminded us the event logs via different formats.

Gallagher C. presented similar work related to this thesis in [6]. Forecasting of sales opportunities concerning winning or losing was provided using the model developed in this research. They considered the qualitative and quantitative data from the Salesforce CRM application. Unlike the event logs in this thesis work, the qualitative data features are the business unit, opportunity region, and deal type. Opportunity contract value, the length of the contract and days in each sales stage, are considered as quantitative data features in this implementation. Bayesian classifiers used for training the cases and their model has the classification accuracy of 93 percent. A Certain period of data used for training and the remaining time used for verifying the cases with the results obtained from the training data. This work reminded us the predictive analysis in sales pipeline.

Schonenberg H. [7] presented a proposal of next event based on the historical activity logs for the unfinished cases. Event logs with completed cases considered as the input for this model which are classified based on the frequency of the traces. Incomplete cases then compared with the closed cases with matching sequence traces, based on the matching criteria recommendation is suggested for the next action. Here the sequence matching was enabled including standards to balance the accuracy. In this research, the implementation was provided in ProM as a plug-in. This work is similar to the method chosen except that prescribed model[7] predicts the next event in the sequence.

Zeng S. [8] presented a supervised learning technique and the equivalent outcomes with invoice-to-cash collections as the background. The accumulated feature was developed to track the payment history for each customer. Results point that holding this set of features enhances the prediction accuracy significantly, and it is more valuable in foretelling payment delays for invoices from returning customers than using customer features. It can perceived from the authors observation data, that delinquency time for the invoice payments can be reduced by the recommendations based on the predictions generated. This work reminds with our study with respect to the predictions based on finished cases.

Gróger, C.[9] emphasizes on the invoice outcome prediction based on the metrics that are used to measure the efficacy of the organization. Time was taken to collect invoice was the most commonly used metric, if you can estimate the outcome of an invoice, then

the collection process parameter can be improved by the predicted information. For example, if you can identify the likely unpaid invoices at the time of creation, then it is easy to reduce the time to the collection by actively trying to collect on these invoices. Usually, collections units wait until the invoices are overdue to initiate the actions like sending reminders, making telephone calls to the customers. It is also trivial regarding business perspective to contact the customer who is likely to pay 90 days later than the one who pays earlier. The author synthesizes different cases to formulate the invoice outcome prediction task as the supervised learning problem like classifying the instances into classes, with the various time periods that form the average overdue metric. This work importance of communication related events in the data.

Xing, Z. [10] presents the data-mining technique for the recommendation based process optimization that explores the prescriptive analytics and provides the action-based recommendations which address the current process issues and improves the current process by making the recommendations that are near to the process. The author defines the recommendation generation feature in four steps starting with identifying the data, mining the model generation, mining the model analysis and recommendation processing. Here the decision tree approach is followed to enable the workflow action recommendations. In the Definition of data basis step, the required data is prepared based on the restrictions on attributes and process instances unlike to our thesis context, but it has got some similarity in action recommendations based on process instances rather than on user actions. Recommendations are derived based on the completed instances data. In the second set, the mining model generation relies on the decision tree analysis. Actions recommendations, in this case, are based on the height of the tree where maximum height defined for the detailed process. In the third step, Mining model analysis, rules for recommendations are set based on the generated decision tree. Each leaf node represents the rule for the recommendation. The final step is to evaluate the rules to adjust the recommendations, although the recommendation generation is explained based on equipment and machinery example that is broadly different from financial indicators. This literature gave us the importance of decision tree that can be used for decision making.

Xing, Z.[11] represented the approach of making recommendations observed from the data mining task, sequence classification for early predictions. Sequence classification has the broad range of application areas like Finance, Healthcare informatics, information retrieval, genome analysis and intrusion detection. In general sequence classification, every sequence of event is linked to one class label or the full sequence before the classification. In sequence classification, there are three trivial things to be considered, for decision trees or neural networks only vector of features can be considered as input. Secondly, when there are different feature selection methods, transformation can be applied to the sequence into the set of features. Third, as no distinct features are building a simple sequence feature is challenging. Sequence classification[11], distinguished in three broader categories, such as feature-based classification, sequence distance-based classification, and model-based classification. In feature-based classification, the transformation of the sequence to feature vector implemented. In sequence distance-based classification, a distance function is associated to measure the similarity among the sequences which eventually determines the quality of classification whereas for the model-based classification other statistical methods like Hidden Markov Model(HMM) are applied to classify the sequences. This work reminded us different ways to construct the sequence data.

Nguyen, H.[12] explains the use of deviance mining in generating the action recommendations in conjunction to process mining. Business process deviance mining are a group of approaches to identify the unusual behavior of a process, to tell why the process is abnormal. This kind of processes can be of the positive or negative type. Active cases are the examples of the high performance enhanced processes with little execution time, cost and resource usage. Negative case deviance is with low performance and adverse outcomes or compliance violations usually with customer complaints. The input for this approach is a collection of labeled traces. Each label for the trace tells whether the process instance is deviant or normal. A classifier constructed takes the input as the traces and outputs the relevant class name. Some observations are presented based on the previous work that is related to software defect handling process, discriminative pattern mining process, frequent pattern mining. All these methods result in distinguishing the positive and negative outcomes.scientific tools used by the authors are Rapid-Miner, MS Access database with vector spaces that has extracted features. For Tandem repeats and the alphabet, variants were derived using ProM plug-in. The accuracy of the data was calculated in folds based on five-fold cross-validation with 80 percent training data and remaining for test purposes. The different interval of results showcased with methods like the decision tree, k-NM, Neural networks. This literature reminded us the model training with different methods for the event based data.

Dumas, M.[13] presented the business process monitoring architecture that highlights the architecture of the firm process solution that relies on both predictive monitoring and deviance mining. Both the methods takes logs with process execution traces and set of business principles. Business principle is the condition evaluated at the completion of a process with yes or no. Business conditions are, for example, Before paying every invoice should be approved or after all the documents submitted, every insurance claim should get resolved within two weeks. When the input logs are ready with business constraints or conditions, deviance mining diagnoses the unusual behavior of the process with positive and adverse outcomes which enables the process improvement circumstances. Once the results generated from the deviance mining approach, predictive monitoring applies the recommendations for process operators while executing the case. These recommendations prescribe the user about the probability of the failure of the particular instance to obey the compliance rules or performance conditions. Predictive monitoring can raise the alerts to the business when some business actions get violated. Whenever there are inadmissible deviations, these business process support systems raise flags by acting as a monitoring and recommendation system. This paper helpful us in identifying and analyzing the positive and negative cases with respect to features.

Liang, Y.[14] researched on failure prediction using activity logs obtained from IBM BlueGene/L. Based on the event logs they developed a prediction model for real failure data. Their prediction model starts with separation of time into determined periods and then attempting to predict if there were collapsed events in each interlude based on the activity characteristics of the earlier interims. Their predictive design spouted two principal difficulties, which were feature picking and classification. Among the different regions of features, authors selected the features based on the characteristics of failure. Authors developed a model based on customized nearest neighbor classifier and compared its performance with classification tools RIPPER(a rule based classifier) and SVM(Support

Vector Machines). They finalized to use nearest neighbor prediction model to optimize the system failure tolerance as it has the best scores. This literature reminded us about the feature selection.

Pika, A, [15] presented a predictive approach for forecasting the deadline overruns which determines whether a case will reach its deadline or not. In the first step, they defined the "Process Risk Indicators" such as abnormal event execution time, abnormal waiting periods, recurring activity repetitions, rarely performed activities, multiple resources involvement. In the second step, they used the statistical methods to recognize the appearance in activity logs by comparing the risk indicators and finding the likelihood of the case being deviant. Insurance company data utilized in the experiments conducted in this paper. This paper reminds our research with respect to the different methods available via event logs.

Pika, A. [16] similar to the above work [15] presented "Process Risk Indicators" to illustrate the prominence of risk indicators applied to forecast the process lags using event logs was presented. The prescribed method learns from the previous process behavior entered in event logs. Three step approach was followed, starting with defining the "Process risk indicators"; in step two, configuring the risk indicators and in step three recognizing the appearance of risk indicators in the current running cases. Classification of the data considered in two splits with maximum records in the training sets. Precision and recall were used to validate the delays occurred, and the implementation in ProM tool was provided as a plug-in. This work reminds with our study with respect ot the classification of the data.

Song, J [17] presented "Behavior pattern mining" considered as subfield to process mining was investigated with the definition of major technologies, data formats, and related tools in this paper. Behavior pattern mining combines the four stages such as events registering, data preprocessing, mining and discovering and the characteristics vary from choosing the most frequent sequence of users, average throughput time of the users, the total duration of the path, mean service time for a particular duty, etc. Authors emphasized on applying the mining algorithms using the ProM tool such as Heuristic miner, Alpha, and Genetic algorithms. We use this work in terms of selection of specific feature(total duration of a case).

Van der Aalst.[18] presented an approach using event logs in combination with compliant "utility" and "compatibility" concepts, a prototype developed to serve as a proof-of-concept applying the ProM plug-in and tested on real-life activity logs. Through the ProM add-ons, authors investigated the procedures to produce advanced event logs to render the proof at the case level of how processes could be gained by shifting origin times of actions and by designating a distinct resource to operate on an event. Authors developed the compound visualizations that call attention to the shifts in periods and alterations in resource distribution linking two harmonious event logs which are used for process improvement inside the company. This paper reminded us in analyzing the importance of resource in event logs.

Bose, R. P. [19] presented a ubiquitous structure for identifying signatures that can be used to describe or forecast the state of noticed and unnoticed traces. The signatures

identified were the discriminative patterns to determine the likely and unlikely actions which help to develop the practices and supervisory processes. Authors followed a systematic approach for the signature discovery starting with the class labeling of the event logs where clustering and classification techniques such as k-means clustering and SVM. In the second step, the features were extracted and selected based on the simple filtering techniques such as principal component analysis. In the third step, patterns were identified to discriminate the faulty cases using decision tree approach and association rule mining. In the final stage evaluation of the signatures, patterns using the goodness of the measures was made. A case study of Malfunctioning of X-ray machines was used as the input data and the implementation was enabled as a plugin to ProM. This work reminds our study with respect to the evaluation of different methods for choosing the optimal model.

3 Background

3.1 Process Mining

Process mining[20] is used to identify, monitor and improve the real-time processes from event logs. Process mining techniques used in the recent commercial products for the process improvement. Process mining in conjunction with data analysis techniques provides an innovative pathway to decision making. Process mining fills the gap between the business process management and data mining regarding business perspective. Execution of business processes based on the operational data from Enterprise resource planning(ERP) and workflow management systems. With the increase in the means, there is a tremendous increase in data in business processes, like operational data, procurement, sales, human resources, logistics, etc. The improved data integration with the information systems and application of process mining does not only provide the aid to enhance the modern information system with efficiency and effectiveness but also contributes new possibilities for data analysis. This data generated by continuous integration of process mining and analysis approach improves the business decisions.

An event log typically consists of timestamps, performer, event id, case id, and the activity.

Process mining[20] consists of event logs, process models. An event log is an ordered set of recorded events of the business activity. A process model instance is the execution of the single process instance with a group of events with a unique identifier. A trace represents the succession of the recorded events in a case. Based on the event logs process models can be extracted which reveal the blueprint of the business process.

An event log typically consists of event ids, timestamps, performers, and the activities. Table 1 represents the event log format [21].

Case ID	Event ID	Timestamp	Activity	Role
1	1	22-05-2006 13:23:33	Activity-A	User-1
	2	13-06-2006 14:23:33	Activity-B	User-2
	3	15-06-2006 14:23:33	Activity-C	User-3
2	4	28-06-2006 09:23:33	Activity-A	User-1
	5	29-06-2006 14:03:33	Activity-D	User-4
	6	30-06-2006 14:23:33	Activity-E	User-5

Table 1: Event logs data structure

- Case ID: Every event in the event logs has the unique identifier to identify the trace. This feature can be the name of the company, the identification number of a group, system-generated identification number.
- Event ID: Application generated identification number of event logs, every record in the event logs has a unique identifier.
- Timestamp: Activity execution time recorded with a time stamp containing the day, month, the year with hours, minutes, seconds and milliseconds. This feature is used to identify the process duration, user activity durations, and activity durations.

- Activity: Every activity in the business process is captured here, which typically consists of user actions, application executions, etc.
- Role: The person who is responsible for the particular event action recorded here. The person here is the employee of the company, client, end user, etc.

The primary process developed for process mining starts with the data extraction from the relevant information source using efficient extraction methods to cope with the amounts of data. Once the data got extracted, filtering of data should be done to prevent some errors and anomalies, followed by loading the data into the process mining software. The process instances can get truncated when the data extracted from the source system. It is trivial to check with the events with the different method as the event is not limited to the single process. Such filtering should be done deliberately to prevent the truncated process instances. In the final step, mining to generate the relationships within the logs and reconstruction of the process model that aims at the discovery of unknown processes. This analysis also gives results for additional objectives like process optimization, organizational facets, and compliance analysis.

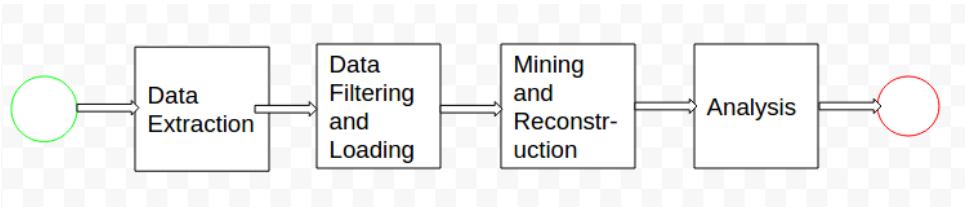

Figure 2: Process Mining Steps [20]

The major areas of application areas include discovery, conformance checking, and extension. In discovery, process models are extracted from the events logs for analysis purposes. Compliance verification takes event logs as the input and makes a comparison with the ideal model to verify whether the design is confirmed or in abnormal behavior. Compliance checking checks whether the internal and external rules followed or not. Operating algorithms in process mining categorized into deterministic mining algorithms, heuristic mining algorithms, and generic mining algorithms. These algorithms determine the creation of process models and its characteristics. Different tools used in the process mining environment, the leading academic and open source tool is ProM, which supports different mining algorithms. Disco is another industrial process mining tool that provides the easy usability options with the unified method for filter and loading of event logs.

3.2 Lead-to-contract process

We used CRM application related data for the experiments conducted in this study. In this section we provide the background about the CRM and Odoo application used in our study.

3.2.1 Customer Relationship Management

Customer Relationship Management(CRM) [22], is the union of the process, people, and technology that investigates to figure out the customer . CRM systems are the part of enterprise information systems which focuses on establishing an organization with customer oriented. CRM[23] systems have emerged in business informatics in customer-focused processes. It is a unified path to organizing customer relations based on customer retention and relationship advancement. Organizations that implement the CRM will profit regarding customer loyalty and long run benefits. CRM successfulness depends on the approach of the team, which is ambiguous to many companies because they don't completely understand the organization-wide, cross-functional, customer- centric business process re-engineering. A balanced way of people, process, and technology, contributes to a successful CRM.

CRM[24][25] technology applications associate both front office and back office activities to bring harmony in between sales functions(sales, marketing, customer service) and financial functions(finance, logistics, and human resources) with companies customer touch points. A companies contact points can incorporate the sales, Internet, e-mail, telephone marketing, direct emails, advertising, call centers, pagers, fax, kiosks, and stores. In few organizations, CRM is just a technology driven software solution to bridge the gap between the sales and marketing to boost target exercises. Other teams deal with CRM for communication with the customer, an exclusive responsibility for sales, marketing and call center divisions. The CRM game plan influences the sales, marketing, customer service, operations, human resources, finance and information technology.

CRM actions[25] have resulted in higher revenues and increased profitability which is, in turn, bringing competitiveness for many companies. Companies like SAP, Oracle, PeopleSoft, SAS, Siebel, Clarify and other companies are competing to showcase new edge CRM[23] applications to the organizations. Advancement in the enterprise software technology and advanced supporting tools enabled with CRM software automates the process of monitoring the customer contacts and helpful in predictions. Technology that monitors and analyzes the customer patterns broadens companies to figure out the best customers and spotlight marketing exercises to reward the frequent buyers. A better understanding of the current customers acknowledges businesses to connect, answer and communicate more intensely to optimize the sales .

The CRM[26] system allows us to

- Identify all the leads coming from different sources like the website, contact forms, etc. and drive them to real opportunities.
- Collects the data related to the customer such as address, email, preferences, etc. to organize the company address book.

- Access all the information, documents, messages connected to leads and opportunities in one common place.
- Maximize the productivity of the sales team by guiding them to be well formulated.
- Helpful in delivering the exact sales forecast by tracking the sales pipeline by stage and timely basis in the form of reports and dashboards.

3.2.2 Odoo CRM

ERP[27] solutions are one of the fastest growing segments in the current world. Odoo also earlier known as OpenERP is a platform for different applications such as CRM, Human Resources, Accounting, Finance, etc. all organized under one platform. Odoo CRM is an open source environment based on the cloud platform. Odoo[26] is contributing its areas in ERP segments; it has made a significant impact on the learning about creating value for the business around the world with its array of features at disposal. Development of Odoo platform is done using three basic technologies which are Python, JavaScript, and XML-RPC.

Odoo CRM[27] is a successful CRM system handling the sales pipeline in an efficient manner. In the following section, we define precisely different terms and application perspectives in Odoo CRM. Lead[28] is the initial contact that can produce business opportunity. Leads contacted via the contact forms on the website; phone calls received from the prospect, business cards, the list of contacts to make calls, etc. Leads in Odoo consists of a quick information storage form which stores basic contact information about the customer such as customer contact details, description, relevant notes and other necessary information required for the business. Leads can also get stored via the database in case of mass leads coming from different sources. All the communication such as sending emails, phone calls related to leads can directly access from this section. Lead can transform to opportunity from the same page at any point of time.

An opportunity[25] is a qualified lead to be analyzed through the companies sales funnel. The sales pipeline in Odoo is organized using the simple and powerful tool in Kanban view. This view enables you the freedom to drag and drop opportunities from one step to the other and to get immediately visible knowledge about subsequent actions, new messages, best opportunities and expected revenues. Following operations can be done in opportunity phase[26].

- Convert a lead to opportunity from the first form.
- Customization of opportunity's Kanban view according to the companies sales funnel.
- When a lead converted to an opportunity, the contact is set as Customer automatically.
- Transition of stages directly from the Kanban view.
- Define and modify the expected revenue, expected close date and stage success ratio based on the business model.

- Option to add a next action and course which is evident in the Kanban view. Apply some tickets to the opportunity, e.g. sections.
- Plan and scheduling the customer meetings. Enabling the Geo-tagging and synchronizing calendar events with Google calendar.
- Communicate with the prospects using chatter via text messages

Sales activities in Odoo structured in a systematic way such that the sales documents can be grouped based on several region, departments, and activities. Allows to assign leads automatically to the correct people in the sales teams. New sales teams can be created and modified including adding the team members, customizing the Kanban view for the sales team just created. Statistics can be drawn for the Kanban board using the system dashboards and reporting section. Apart from these core features Odoo also provides many other add-on features by plug-ins.

This research work is conducted based on a SaaS based company. They formulated different stages in Odoo CRM based on their business needs and structure. Table 2 represents different stages and its probability ratios in Odoo CRM.Odoo sales pipeline has eight stages which could be time-consuming sometimes. The process generally starts with the collection of leads from different sources. Once the leads get captured they are assigned to different sales people(users) inside the organization to work in close collaboration with the leads to make the initial contacts. After the particular time, if the customer is willing or showing interest toward the product those leads can get converted to opportunities and in some cases some clients drop down in this stage, they can be marked as dead. Once the opportunity step is confirmed, they can be managed in different stages according to the business requirements. The semi-final stage is to identify the qualified opportunities, the one's which are qualified can be marked for agreement and the ones not qualifies ends there itself. After the approval stage if the customer is willing to buy the product they can mark as won deals and deal can be closed, and the invoice gets generated at the end. In the other case when the customer withdraws at the agreement stage can be marked as lost, here a follow-up call should be made to confirm, these opportunities can also be again marked as leads.

	Stage Name	Probability	Type
1	Qualification	5	opportunity
2	Kohtumine/prese(meeting)	25	opportunity
3	Pakkumine(offer)	17	opportunity
4	Leping(Agreement)	50	opportunity
5	Lost	0	opportunity
6	Won	100	opportunity
7	Pole hetkel turul(Currently on the market)	0	lead
8	Ei kvalifitseeru(Does not qualify)	0	lead

Table 2: Description of stages in the CRM pipeline

- Qualification: This stage is the initial stage in the opportunity level. When the sales person converts from lead to opportunity, those opportunities fall under this category. This step mentioned as the qualified leads which need further work to

close the deal. Leads get converted to qualification stage by any salesperson on the sales team. According to the companies business criteria, when a lead falls under this category has a success ratio of 5

- Kohtumine/prese(Meeting stage) : This stage is also known as the meeting stage. In this stage mostly the meetings with the first opportunities are scheduled with prior intimation to make discussions about the deal. Opportunities are at this juncture has 25% probability of closing the deal.
- Pakkumine(offer stage): Opportunities that fall under this stage reached the proposal stage in the sales pipeline. This stage has the 17% probability of closing the deal.
- Leping (Agreement stage) : This step is considered as the significant stage in the sales pipeline as most of the deals decided at this juncture. This stage has a success ratio of 50% probability of winning the case.

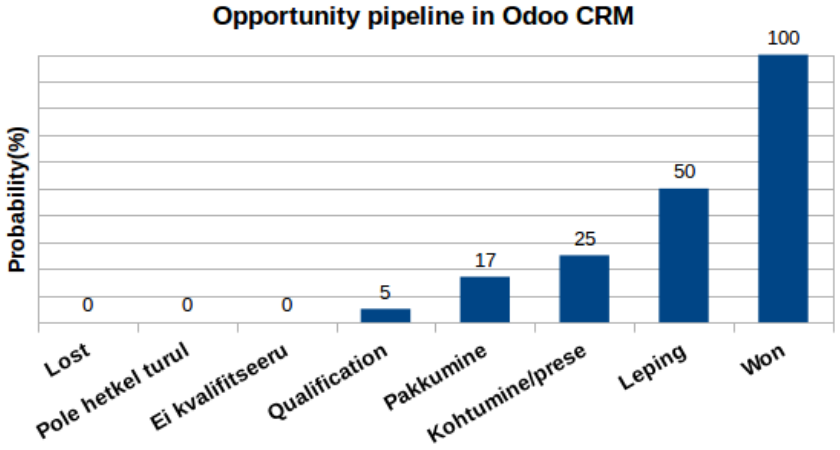

Figure 3: Distribution of stages in Odoo CRM

- Won: Closed deals are placed at this stage which means they successfully made the deal and signed the contract with the company. This stage probability ratio is 100% success.
- Lost: Potential deals lost at any stage in the sales pipeline can be converted to a lost opportunity which has 0% probability of success ratio.
- Ei kvalifitseeru (Does not qualify): The cases which does not qualify in the sales pipeline converted to this plane which has 0% probability of closing the deal. These cases can also eventually get converted to leads.
- Pole hetkel turul (currently on the market): This stage is has a probability of 0% success ratio of closing the deal. The leads which cannot get considered for the sales pipeline gets converted to this stage.

3.2.3 Lead management and Opportunity funnel

Following steps represent the brief explanation about the lead management and sales pipeline in Odoo CRM based on a Saas company,

1) Sales manager loads into CRM cold leads (the manager filters the list from the database with respect to the specific attributes of companies. (e.g. activity field, annual revenue, number of employees)

2) A sales representative marks first leads to qualify for an opportunity based on the outstanding features to be qualified.

   2a) After getting a lead qualified, the aim is to schedule a meeting for a presentation (follow step 3b).

3) For other cases (which is more common), the sales representative will call the customer for qualification.

   3a) If the person cannot get the contact with a proper person or no one answers the phone, then the person schedules another call at the CRM and sends an e-mail to the customer.

   3b) If the right person contacted then a meeting is scheduled, and the lead marked as "Presentation" without marking it "Qualified" (if this is the first iteration), and a meeting scheduled at the calendar.

   3c) if the person is not interested in the meeting, then the lead/opportunity marked as "Lost" and summary of notes is provided to the opportunity description for future use.

   3d) if the person cannot immediately decide when to have the meeting, then the lead is marked as "Qualified," a follow-up call is scheduled and the cycle proceeds from 3b until the sales person are willing to do so.

4) The sales representative decides the opportunity decision in based on different paradigms.

   4a) if the customer shows positive towards the business then in the meeting then the opportunity is marked as "Won."

   4b) else an offer sent to the opportunity, and the stage of the opportunity is set to "Pakkumine" and a follow-up call or a meeting is scheduled; there may be several follow-up calls before the stage of the opportunity is changed at this moment.

   4c) if the offer, presented at the meeting, was all right but the client did not want to proceed instantly, then set stage to "Leping".

After sending out an offer, the sales process slows down due to less pressure to the opportunity, and this often leads to lost opportunities. Lost opportunities are potential new leads and opportunities. If the first follow-up is 4-5 days after the day when an offer was made then the opportunity/lead gets significantly "colder".

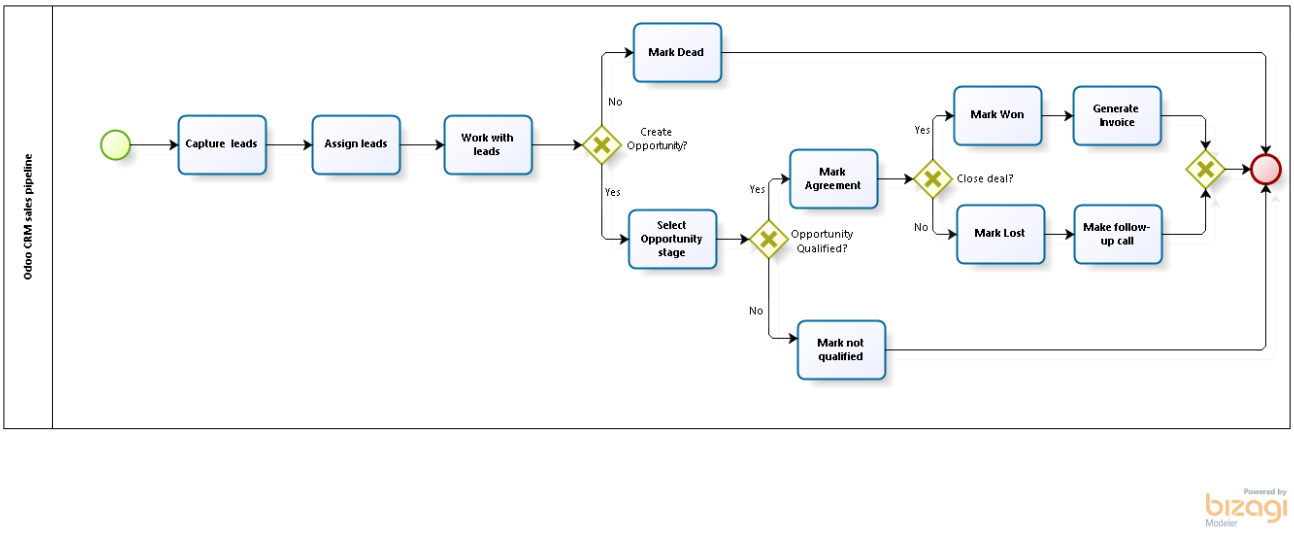

Figure 4: Process Model representing Odoo sales work-flow

Based on the leads management and opportunity funnel, a process model is generated which can be seen in Figure 4. Odoo sales pipeline starts with capturing of leads. Capturing the leads is an ongoing process. Captured leads are assigned to sales representatives in the company then the specified steps are followed to decide won/lost opportunities at the end of the process.

3.3 Tool description

In this section, we define the tools used in our research work and their relativeness in the context of implementation.

3.3.1 R

We used R[29] in this thesis work for data preprocessing and predictive model construction. R is a programming language for statistical software development and data analysis developed by AT&T Bell Laboratories. R[30] is maintained and distributed by reputed software statisticians and scientists from universities and software industries. R is the prominent tool used for the research in statistical analysis, social science, economics and enterprise sector. R contributes to the data analysis in wide areas using the classical techniques such as linear and non-linear modeling, time-series analysis, clustering, predictive models using different algorithms and others. R[31] is similar features with object-oriented programming languages. R has embedded abilities for extra functionalities via user built packages which give access to specialized statistical techniques, advanced graphics, and export/import applications on external data formats. R is platform independent which works in Unix, Linux, Windows, and Mac OS[32].

We used R extensively for data preprocessing and method implementation including different machine learning methods such as Random Forest, Naive Bayes, eXtreme Gradient Boosting, Single C5.0 Ruleset, Penalized Discriminant Analysis, Partial least squares and Support vector machines.

- Caret[33] (classification and regression training): The caret package used in solving classification and regression problems, which has different tools to develop the predictive models based on the available models in R. This package is mainly helpful in classifying the models and tuning with multiple modeling techniques. It is also useful in mapping the variable importance from the model for feature selection, model visualizations, data preprocessing for training set data.

- Random Forest(RF): Random forests[33] are the ensemble based methods for regression, classification, and some more additional methods. Random forest constructs the decision trees at training time and outputs the corresponding predicted class in case of regression problem and mode of the classes in case of classification problems of the individual trees. Random forests grow multiple classification trees, for classification of new vectors, it considers the votes for the class and selects the class which has a maximum number of votes. Random forest classifiers trained for larger cases and processing times are minimal.

- Naive Bayes(NB): Naive Bayes[33] classifiers comprise a group of naive probabilistic classifiers based on implementing Bayes theorem by robust independent ideas between the features. They are profoundly scalable classifier type of methods, which require some parameters extended in the number of features. For few representations of probability models, Naive Bayes can be prepared quite predominantly in a supervised learning environment. In various functional applications, feature judgment for Naive Bayes classification practices the means of highest likelihood.

- eXtreme Gradient Boosting (EGB):[34] is a productive package with a tree learning algorithm and a linear model solver. Users can define their objectives of the data as

this package is extensible. The primary purpose of EGB is to train a booster that predicts the prediction to the model. Principal features of EGB include: Sparsity, customization, input various types, can perform parallel computation on Windows and Linux with openmp.

- Single C5.0 Ruleset (C5.0 Rules):[35] This method is an extension of the C4.5 classification algorithms. The design model can assume the structure of a complete decision tree or a group of rules or boosted variants of both. When utilizing the specifications method, factors, and additional classes are saved (i.e. duplicate features created automatically). This distinct model manages nonnumeric data of few varieties (such as characters, factors, and normalized data).

- Penalized Discriminant Analysis (PDA):[36] is a technique which is primarily useful in extracting and interpreting discriminant functions and to design a clear understanding of the scientific data. The predominant usage of this technique appears to be in the speech and image recognition where there is a huge volume of correlated inputs. In the context of medical needs, there is an immense need to have a well-organized prediction performance and the high degree of interpreting ability on the smears and mammograms where PDA is a sagacious technique to approach.

- Partial least squares (PLS):[37] used in fabricating predictive models especially in the industrial applications of a flexible modeling technique. When there are a vast number of factors or variables, it is extremely challenging to obtain an unambiguous predictive model. In such cases PLS plays a key role in designing an accurate predictive model for the data available. The principle of PLS lies in the indirect extraction of latent variables from sampled factors and responses. The extracted factors are then used to construct the predictions for the replies. It mainly emphasizes on the collinear factors in predicting the models. This method includes three main techniques: Principal Components Regression (PCR), Maximum Redundancy Analysis (MRA), and Partial Least Squares.

- Support vector machines (SVM): This method[33][38] developed by developed by Cortes & Vapnik (1995) is the notion of decision planes that connected decision frontiers. Decision plane separates between the set of things having changed group memberships. SVMs are applicable in different areas such as classification, regression, and outliers discovery with an intuitive model design. In R, this method is available via 'caret' and 'e1071' packages which offer different classification and regression features such as formula interface, sigmoid kernels, linear, polynomial, radial basis function, and k-fold cross validation.

3.3.2 Disco

We used Disco Fluxicon[41] for generating the process models from event logs which give the overview of the process. Disco is a process mining tool which is efficient in importing and exporting of an event log in a fast and easy methods. In process mining field, Disco is the user-friendly tool which users can access it without any prior knowledge of the application. This tool handles large datasets which are suitable for event logs. This tool is most favorable for the event logs data with more complexity and complications which can be drawn from process model to avoid ambiguity. Another benefit in using Disco is that filtering of data can be done in depth depending on the criteria. Apart from creating

process models from event logs, Disco can also create an animation from event logs which can precisely detect the potential problems within the process model or an event log. Another advantage of using Disco is that it allows various input formats of data such as CSV, XES, and MXML. This tool also supports converting event logs from one form to the other format.

3.3.3 ProM

In our research work, we used ProM[40][39] for visualizing and investigating the event logs data sets in new ways for understanding the dataset. ProM is the Open source process mining tool which handles event logs to generate different graphs, BPMN models, C-nets, Yawl diagrams, etc. ProM is efficient in investigating the abnormalities and deviations probably prevailing in data by comparison of actually collected data with a standard master model. Moreover, ProM allows reconstruction of resulting graphs with many available models (Petri-nets, Process models, etc.,). ProM supports integrated process mining, various formats supported by different processes, algorithms and the large variety of languages such as Petri-nets, Social Networks, etc., can be added. Another advantage of using ProM is it supports a plug-in based environment where researchers and developers can contribute to add extra functionalities via plug-ins.

4 Data

4.1 Structure

This thesis work is based on the data from Odoo CRM explained in the background section, from the application via the database the raw logs are extracted from different sources and preprocessed to obtain the event logs necessary for this work. Event logs in our implementation consists of different timestamps (day, month, year, hour, minutes and seconds), events, users, case identification, labels, company names, etc.

Create Date	User Id	Case Id	Body	Record Name
2015-12-01 08:35:30	5	82	Opportunity has been convertedto the quotation SO006.	xxxx
2015-11-30 10:07:59	5	82	Stage changed Stage: Pakkumine → Leping	xxxx
2015-11-30 10:08:45	5	82	Expected Revenue: 4630.0	xxxx
2015-12-01 08:35:01	5	82	Opportunity won Stage: Leping → Won	xxxx
2015-11-05 12:07:39	5	82	Stage changed Stage: New → Negotiation	xxxx

Table 3: Raw data from Odoo CRM

Table 3 shows the extracted raw logs from Odoo CRM for the processing of event logs. Only a few notable features are displayed in Table 3. Only data related to case number '82' are presented here.

- Create Date: Unordered list of timestamps during activity start time were recorded here.
- User Id: User responsible for the activity. Used in event logs directly without any processing.
- Case Id: Instance identifiers are recorded in this feature. No additional processing is done while preparing event logs for this feature.
- Body: This feature typically contains the activities carried out in Odoo. As Odoo includes different applications, data filtration unique to CRM was implemented for event logs generation.
- Record Name: Customer names are observed in the record names.

Table 4 shows the event logs snapshot based on different sources which are explained in Section 4.2

Log ID	Timestamp	User ID	Case ID	Case Name	Event	Label
1	2015-11-05 12:07:39	5	82	xxx	Stage: New → Negotiation	0
2	2015-11-05 12:07:39	5	82	xxx	Phonecall created	0
3	2015-11-30 10:07:59	5	82	xxx	Stage: Pakkumine → Leping	0
4	2015-12-01 08:35:01	5	82	xxx	Stage: Leping → Won	Opportunity Won
5	2016-01-12 07:07:20	11	295	xxx	Stage: Qualification → Kohtumine/prese	0
6	2016-01-18 13:00:49	11	295	xxx	Stage: Kohtumine/prese → Won	Opportunity Won

Table 4: Processed event logs from Odoo CRM

- Log ID: Unique identifier of the records in the event logs.
- Timestamp: Activity start time is recorded with a time stamp containing the day, month, the year with hours, minutes, seconds and milliseconds. This feature is used to identify the process duration, user activity durations, and activity durations. The dataset used in this thesis has the data from time line November 2015 to May 2016.

- User ID: Identifier of the user responsible for the particular event action is recorded here. All the members in the company like sales representatives, managers, etc. fall under this category.
- Case ID: Identifier of a group of events belonging to the same sequence which is used for construction of sequence-based data modeling which is explained in Section 4.3
- Case Name: We use customer names as the case names.
- Event: Type of event such as "phone-call created", "Meeting scheduled" etc. which can be seen in Figure 5
- Label Represents the closed cases which are used to identify positive and negative cases.

4.1.1 Events and labels

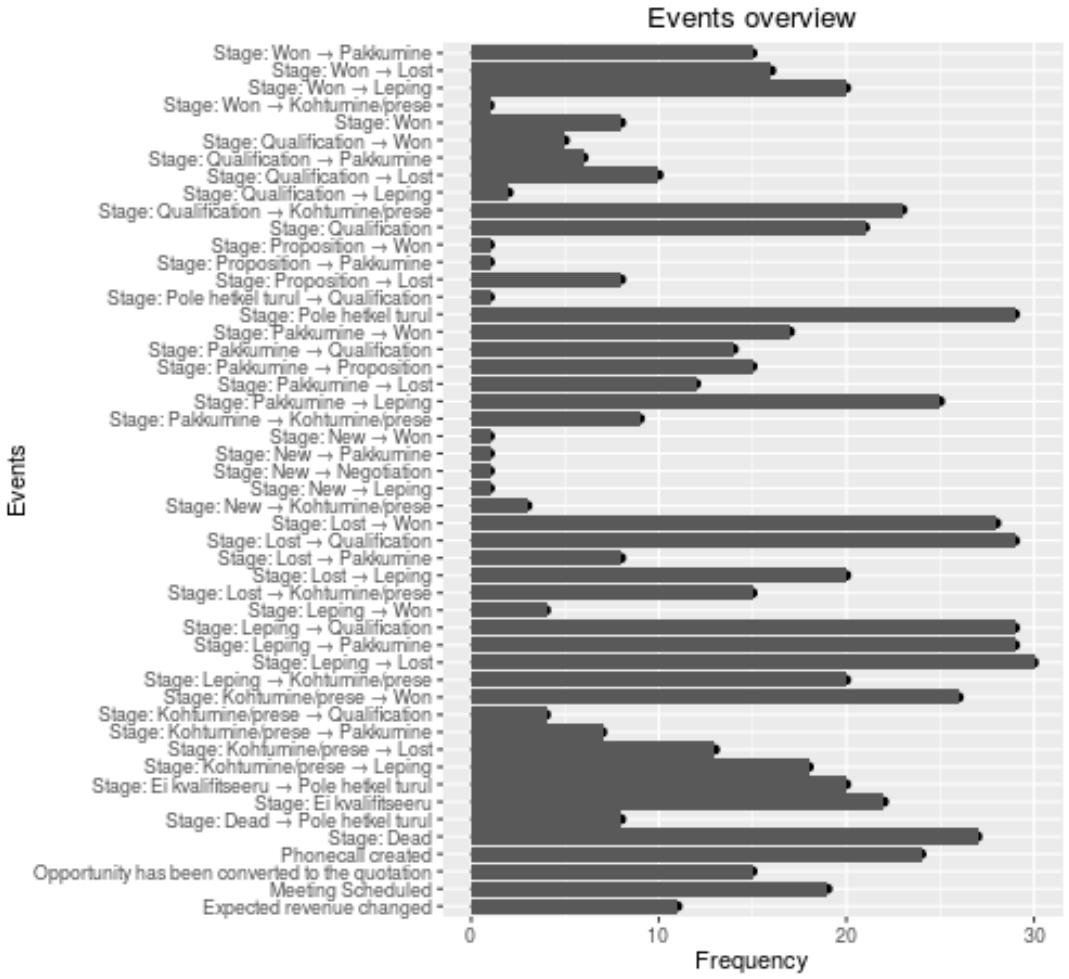

Figure 5: Overview of the events in the event logs

The event logs extracted from raw logs from various sources, has 49 event types, which are displayed in Figure 5 are specific to stage transitions, phone calls, meetings,

etc. Every event has a label specifying the end of the event. For example, if the stage ended up with a negative label, then the label is "Opportunity lost". The label indicates the finished events of the sequence (won or lost). Labels represent the closed cases which have two types of values they are 'Opportunity won', 'Opportunity lost' while the others are marked with "0" which is also added to different data formats prepared further after pre-processing of data.

All the stages displayed in events constituting to positive and negative labels are represented with the labels separately. Transitions observed in the events are represented by arrow sign. Few events are discarded from the event logs to bring the consistency while training data. Events with finished labels such as "Stage: Won", "Stage:leping -> Won", "Stage: Dead", etc. are removed to balance the dataset with only the event steps not containing the last(finished) event in the sequence. Processed event logs at this stage are ready for the preparation of target datasets used for model training explained in Section 4.3. Event types are explained as follows,

- The state change of events: The change of state of events from one stage to the other stage falls under this category. Among the 49 event types shown in Figure 5, 45 event types belong to stage transitions represented with a "Stage" tag, for example, "Stage: Leping → Won".
- "Phonecall created": This stage indicates the communication in the sales funnel. Whenever a phone call is performed in Odoo CRM, this activity is recorded in the event logs.
- "Opportunity converted to the quotation": Only a few cases in our dataset are marked with this label which signifies the potential sale. This label is observed after the opportunity is won.
- "Meeting Scheduled": Whenever a meeting is scheduled with the customer in the Odoo opportunity funnel, this name is recorded in the events field in event logs.
- "Expected revenue changed": Adding or modification of estimated revenues of the opportunity inside the Odoo sales funnel are recorded with this name.

	Name	Total #
1	Event types	49
2	Case labels	2
3	Event logs	6904
4	Total cases	1038
5	Finished cases	776
6	Unfinished cases	262
7	Positive cases	345
8	Negative cases	431

Table 5: Event log characteristics

Table 5 represents the event log characteristics that has 8203 total number of event log records after successful preprocessing of raw data. With the help of the label, the cases in the dataset are 1038 among them, 776 cases are finished cases, and the remaining

unfinished cases are about 262. Finished cases end up with positive and negative labels. Only finished cases are considered for training methods described in Section 5 whereas unfinished cases are considered for providing the predictions based on the results obtained using finished ones.

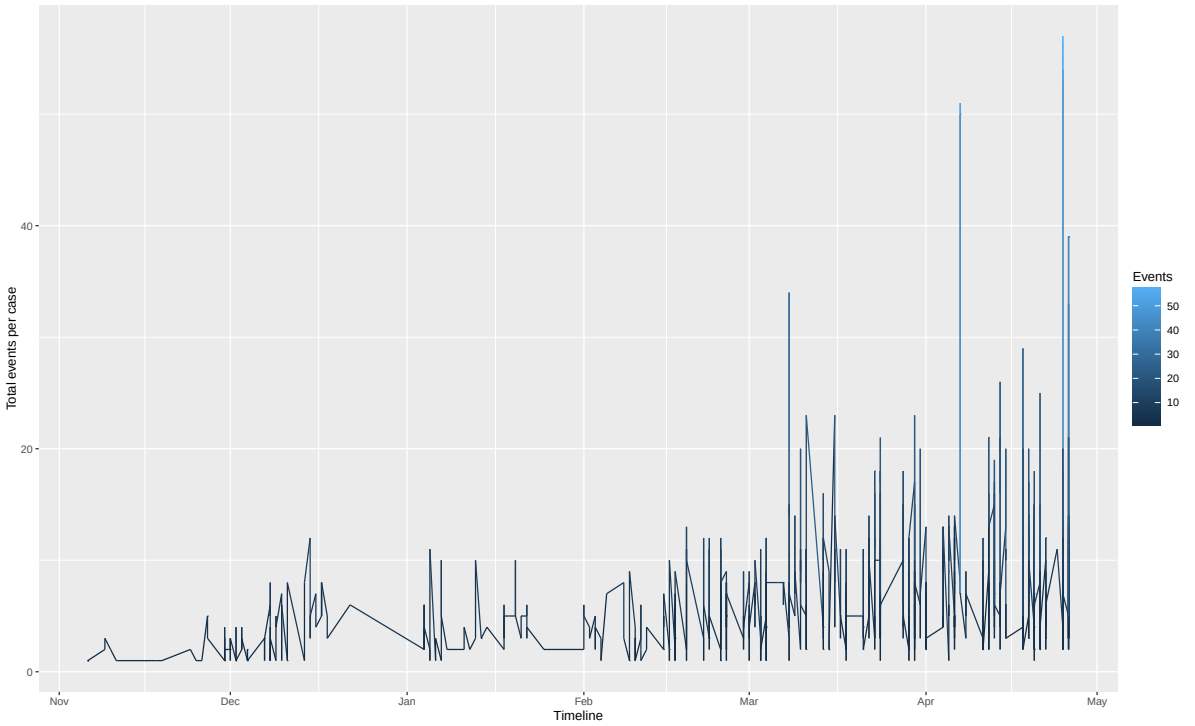

Figure 6: Overview of the events with time line

Figure 6 represents the overview of the events in the sequence of the timeline. The entire schedule showcased in Figure 6 signifies that the events activity started in November 2015 and ended in April 2016. Initially lesser event sequence dimensions observed in the initial months with an increasing trend in the event action length on the timeline. This property of differentiation of event lengths in the periods might be due to the increase in the cases with respect to time. The complete execution of a sequence contains different event transitions and communication with the customer via phone calls and meetings.

Labels convey the finished cases in the binary form, i.e., if the opportunity is lost it is represented with "0" and if the opportunity is won, represented with "1". This feature separates the completed cases from the running cases. Labels are useful for training the model which substantiates as a predicate.

4.1.2 Users

Figure 7 describe the structure of the users and their frequency of the activity in the sales pipeline. Users are the process operators who are the respective sales representatives, managers and other staff in the company responsible for processing the CRM application. In the event logs, all the users perform certain actions which are denoted by

the ID codes, and respective timestamps. Every user ID associated with the user's profile details such as the name, role, etc. is estimated as the important feature for classification.

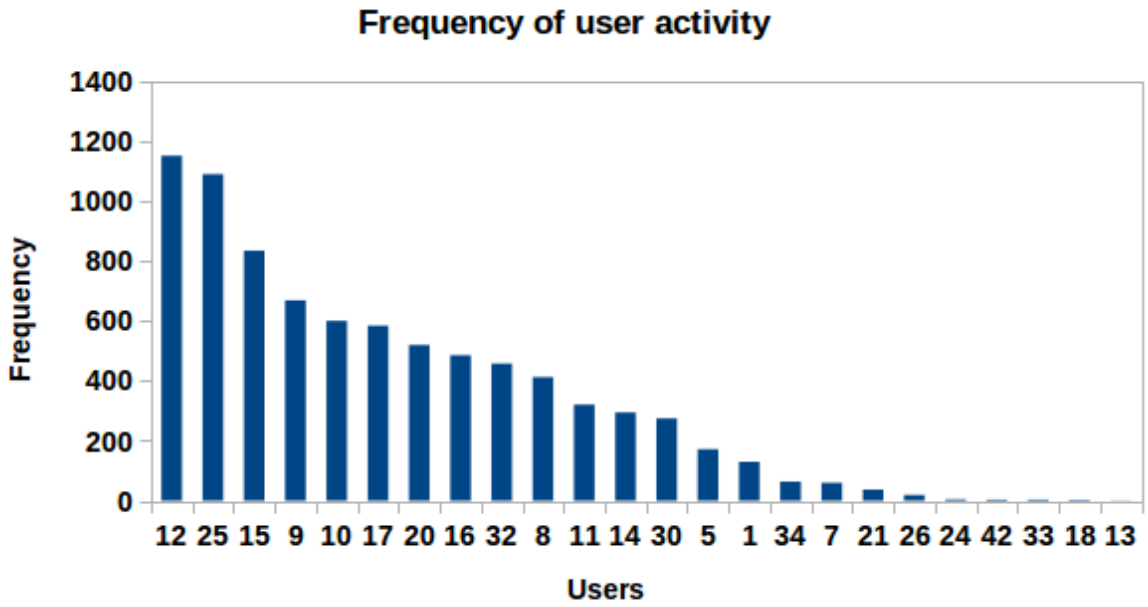

Figure 7: Overview of the users in event logs

Figure 7 represents the activity of the resources in the event logs. More active users can be found on the left side of the graph and least active users to the right of the chart. The active users are the sales representatives of the company handling the day to day opportunities in the sales funnel whereas the inactive users are not directly responsible for managing the opportunities in the sales pipeline, these users might be the leading roles in the company.

4.2 Sources

The event logs data used in our research work is derived from different data sources from Odoo CRM and external datasets. Event logs in Odoo CRM are recorded internally without any external plug-ins, and they are stored in specific database table in PostgreSQL. The data table consists of different large feature set including timestamps, users, actions, model types(specific to applications in Odoo), email addresses of the users, record names and other application specific data. Although there is large feature set, we considered only timestamps, user IDs, user actions and case IDs for the construction of event logs. In Figure 8 data sources are represented with the preprocessing to derive the target data.

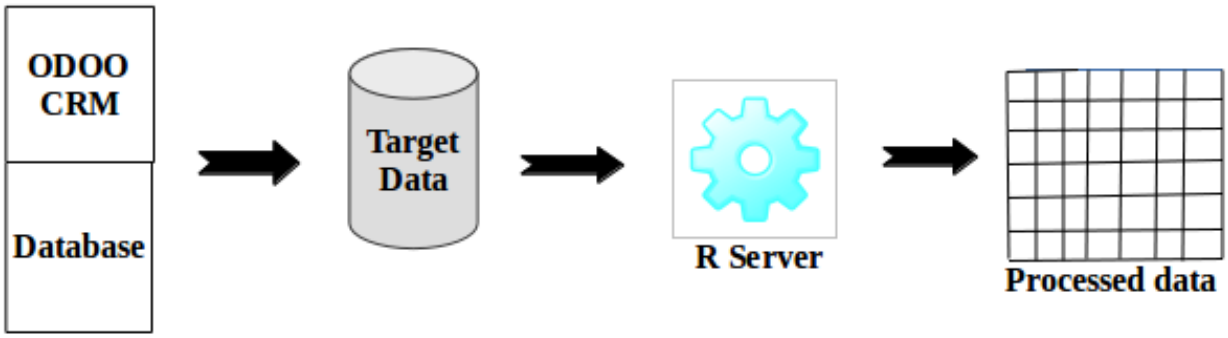

Figure 8: Data sources and preprocessing steps

Raw data has 129870 records which are processed to obtain the target data(event logs) which consists of 6904 records avoiding the data that don't fit in our study. The following is the sample snapshot(first record) from the raw data.

```
"50055 2016-01-21 07:31:16.654409 2016-01-21 07:31:16.654409 NA 12 12 4358
6 720 1453361476.812320947647095.837415771793483-openerp-720-crm.lead@odoo-irtest
crm.lead 0 f 2016-01-21 07:31:16 610 notification Xxxx Xxxx  xxxx@xx.ee  Xxxx X
xxx  xxxx@xx.ee t Stage: Qualification -> Lost     Opportunity lost      "
```

Raw data is in the form of HTML tags which is processed using R as explained in Section 4.3. The raw data displayed contains all the required information for the event logs such as the activities, timestamps, users, etc. Although various other data features were present such as email communications, contact details of the user, application specific settings, we discarded those data as they are not useful in our study.

Table 6 presents the dataset characteristics after processing the data from different sources.From the event logs events signifying the finished activities were removed for the model implementation purpose to increase the consistency in the dataset.

Odoo captures the process activity logs data into "mail_message" dataset which aggregates the data related to the different applications available on the platform such as

Human Resources, Accounting, Logistics Management, CRM, etc. We considered data particular to CRM as we are dealing with the event log analysis of lead-to-contract process (CRM).

	Dataset	Total Records#
1	Raw	129870
2	Event logs	6904
3	Completed cases	776
4	Running cases	262
5	1-fold event cases	371
6	2-fold event cases	284
7	3-fold event cases	264
8	4-fold event cases	173
9	5-fold event cases	145

Table 6: Dataset characteristics

Various other sources of data are used for the construction of data because of lack of required information in the event logs from the database table containing logs. Variable features such as "record name" containing missing values which in turn is bringing inconsistency to the data, to balance this problem we used specific data tables from database which fills this gap. The database tables used for the construction of event logs are 'mail.message', 'res_partner', 'crm_lead' and 'crm_phonecalls'. 'mail_message' data table has the information related to event logs.

Customers ("res_partner") has the information about the customers such as the customer profile details, if the customer is a company then the specific details are available. Lead ("crm_lead") has information related to the opportunity such as different timestamps specifying various actions, financial data (Estimated revenues, Liabilities, etc.), customer names, customer registration codes, etc. Registration codes available from the leads data are used for the construction of the external data as the registration codes are the key connection parameter for the external data.

Phonecalls available via database (crm_phonecall) are used for the enabling the communications data for the event logs. Data such as the timestamp of the phone call, duration of the call, ID code of the customer, etc.Folded event step datasets (1,2,3,4,5) were reduced due to filtering of the finished events from the datasets and few missing data such as external data attributes for some cases.

For implementation purposes external data is added to the event logs for the feature construction. External data has tax debts, vat declarations etc. which is extracted from the external data sources of the company.

4.3 Data pre-processing

In our thesis work, we processed data in various formats for method implementation purposes using R programming. Following are the ways and table formats used in our approach.

- In the initial processing of event logs raw data is taken as input and event logs described in Table 4 is obtained as the outcome of preprocessing. In this phase we processed the raw data which is containing the HTML based encoding in the event feature, where we extracted the required events from the raw dataset. Record names(company names) are added to the missing values to make use of more records available in the data. Moreover in this phase, we screened out the data specific to CRM model as the raw data contains all the data from different models. Noisy data such as duplications, redundant values are omitted for consistency in data. The main preprocessing functions followed at this level are
  - Data extraction: In this step, we extracted the data defined from different sources and loaded into R for preparing the desired data compositions.
  - Data splitting: In this step, we split the event logs into different formats based on sequence-based encoding, user-based encoding, and sequence based encoding.
  - Data re-ordering: In this step, we arranged the data based on the timestamps in increasing order.
  - Removing noisy data: In this step, we removed the noisy data such as finished activities in the event logs which helps to bring consistency while training the model.
- Internal data features such as total duration of the sequence based on start date and end date are considered in days and hours format, total estimated revenue are extracted for the construction of datasets. Durations of the cases in days and hours format are extracted from the event logs directly whereas the estimated revenues are extracted from the database table specific to leads. Additional to this feature vector, total events in the sequence were also added in this dataset. Table 7 showcases the additional processed features.

Case ID	Total events	Start date	End date	Total duration	Estimated Revenue
82	9	05.11.2015 12:07:39	08.12.2015 11:26:17	32 days, 23 hours	2315.0
185	2	06.11.2015 15:36:23	02.12.2015 06:53:20	26 days, 15 hours	0.0
194	2	05.11.2015 06:38:16	27.11.2015 08:22:37	33 days, 1 hour	1320.0
196	6	06.11.2015 08:57:17	05.01.2016 09:36:37	60 days, 15 mins	2230.0
198	8	2015-11-30 15:36:23	18.01.2016 08:50:33	49 days, 18 hours	2750.0

Table 7: Internal data from event logs

- In sequence-based encoding we followed the state of art technique[1]. Every case is classified in terms of sequence based on timestamp in ascending order. For example if a case contains a sequence of 6 events then they are arranged in a sequential order according to the timestamp event-1, event-2, event-3, event-4, event -5, and event-6. So every row consists of the case ID and the sequence of events. Table 8 illustrates the sequence encoding with a snapshot of first few cases from the real dataset.

Case ID	Event-1	Event-2	Event-3	.....	Event-n
82	Phonecall created	Stage: New → Negotiation	Stage: Pakkumine → Leping	.....	Stage: Leping → Won
185	Stage: Qualification → Won	Expected revenue changed	0	.....	0
194	Stage: Qualification	Stage: Qualification → Lost	0	.....	0
196	Stage: Qualification	Stage: Qualification → Pakkumine	Stage: Pakkumine → Leping	.....	Stage: Leping → Won
198	Stage: Qualification	Stage: Qualification → Leping	Phonecall created	.....	Stage: Leping → Lost

Table 8: Sequence-based encoding of events

- User-based sequences similar to event based sequences are represented in Table 9. Users represent the users who is responsible for every action in the sequence matrix. For example if the event-1 is triggered, respective user who is responsible is user-1. All the users responsible for the actions are in the sequence are concluded into a dataset.

- In frequency-based encoding of data, frequency of event occurrence in a sequence is counted as shown in the Table 10. In Table 10, columns names are the events and the rows has the relevant frequency of event occurrence for that case. The events in the matrix are arranged unevenly without following the sequence order seen in Table 8. Rows in the dataset represent the frequency of events labeled in the columns. All the activities available in the event sequences particularized into the columns. Appearance of the activity in the case denoted by its number and absence of the events marked with '0.' For example, in particular, case, if a phone call made, is twice then it is represented by '2.'

- Adding external data features: External data is considered from the external databases of the SaaS company which we mentioned in Section 3.2.3

Case ID	user_1	user_2	user_3	user_4	.....	user_n
82	5	5	5	5	.....	5
185	12	16	0	0	.....	0
194	8	8	0	0	.....	0
197	8	0	0	0	.....	0

Table 9: User-based encoding

Case ID	Expected revenue changed	Meeting scheduled	Phonecall created	.....	Stage: Qualification → Pakkumine
82	0	0	1	.....	0
185	1	0	0	.....	0
194	0	0	0	.....	0
196	0	0	0	.....	1
198	0	0	3	.....	0

Table 10: Frequency-based encoding

Case ID	pr	deg	score	td	tdp	tdd	tdi	tdip	md	decl	age
82	2.79323902187544e-06	3	3	0	0	0	0	0	-1	0	5807
194	2.33446321352797e-06	1	1	0	0	0	0	0	-1	0	5849
198	3.75742658903361e-06	3	3	0	0	0	0	0	0	0	6943
236	1.7981676104202e-06	1	1	0	0	0	0	0	-1	0	6424
249	1.84460115476131e-06	2	3	0	0	0	0	0	-1	0	5472

Table 11: External data features

* pr: Page Rank of a company in the board member network
* deg: Degree of a company in the board member network
* score: The number of problematic companies in the neighborhood of the company in the board member network
* td: Tax debts
* tdp: Tax debts postponed
* tdd: Disputed tax debts
* tdi: Tax debts interests
* tdip: Tax debts interests postponed
* md: The number of non-submitted annual reports
* decl: Declarations of a company in the board member network
* age: Age of a company in the board member network

Based on Table 8 and Table 9, event fold formats are prepared for the method implementation. We focused on preparing the datasets based on n-event matrix approach. In n-fold event approach we consider the event sets and users in the following order with length up to 5.

1-event matrix: First event and first user of the instance in the event logs.

$$\{event-1, user-1\}$$

2-event matrix: Two first events and users of the instance in the event logs.

$$\{event-1, user-1, event-2, user-2\}$$

3-event matrix: Three first events and users of the instance in the event logs.

$$\{event-1, user-1, event-2, user-2, event-3, user-3\}$$

4-event matrix: Four first events and users of the instance in the event logs.

$$\{event-1, user-1, event-2, user-2, event-3, user-3, event-4, user-4\}$$

5-event matrix: Five first events and users of the instance in the event logs.

$$\{event-1, user-1, event-2, user-2, event-3, user-3, event-4, user-4, event-5, user-5\}$$

Case ID	event-1	user-1
82	Phonecall created	5
198	Stage: Qualification	8
249	Stage: Qualification → Pakkumine	8
295	Stage: Qualification → Leping	11
347	Meeting Scheduled	16

Table 12: 1-event matrix

Case ID	event-1	user-1	event-2	user-2
82	Phonecall created	5	Stage: New → Negotiation	5
198	Stage: Qualification	8	Stage: Qualification → Leping	8
249	Stage: Qualification → Pakkumine	8	Stage: Pakkumine → Leping	8
295	Stage: Qualification → Leping	11	Phonecall created	11
347	Stage: Qualification → Kohtumine/prese	16	Meeting Scheduled	16

Table 13: 2-event matrix

Table 12 represents the first event and first user of the cases in the event logs. Table 13 showcases the 2-event matrix representation which has first two events and users of the instance in the event logs. Different feature combinations are added to this dataset to feature importance.

Table 14 represents the 3-event matrix representation which has first three events and users of the sequence in the event logs which again forms six different datasets with the available feature sets.

Case ID	event-1	user-1	event-2	user-2	event-3	user-3
82	Phonecall created	5	Stage: New → Negotiation	5	Stage: Pakkumine → Leping	5
198	Stage: Qualification	8	Stage: Qualification → Leping	8	Phonecall created	8
249	Stage: Qualification → Pakkumine	8	Stage: Pakkumine → Leping	8	Stage: Leping → Qualification	8
295	Stage: Qualification → Leping	11	Phonecall created	11	Stage: Leping → Kohtumine/prese	11
391	Stage: Qualification → Kohtumine/prese	9	Expected revenue changed	20	Stage: Kohtumine/prese → Leping	20

Table 14: 3-event matrix

Table 15 is the 4-event matrix representation which is the batch for firth four events and users of the instance in the event logs. This matrix is the baseline for formation of 6 different datasets.

Case ID	event-1	user-1	event-2	user-2	event-3	user-3	event-4	user-4
198	Stage: Qualification	8	Stage: Qualification → Leping	8	Phonecall created	8	Phonecall created	8
249	Stage: Qualification → Pakkumine	8	Stage: Pakkumine → Leping	8	Stage: Leping → Qualification	8	Phonecall created	8
295	Stage: Qualification → Leping	11	Phonecall created	11	Stage: Leping → Kohtumine/prese	11	Stage: Kohtumine/prese → Qualification	11
391	Stage: Qualification → Kohtumine/prese	9	Expected revenue changed	20	Stage: Kohtumine/prese → Leping	20	Expected revenue changed	10
592	Phonecall created	9	Stage: Qualification → Leping	9	Phonecall created	9	Phonecall created	9

Table 15: 4-event matrix

Table 16 is the 5-event matrix representation which showcases the first five events and users of the instance in the event logs which acts as the baseline for formation of different feature sets based on the available features.

Case ID	event-1	user-1	event-2	user-2	event-3	user-3	event-4	user-4	event-5	user-5
198	Stage: Qualification	8	Stage: Qualification → Leping	8	Phonecall created	8	Phonecall created	8	Phonecall created	8
249	Stage: Qualification → Pakkumine	8	Stage: Pakkumine → Leping	8	Stage: Leping → Qualification	8	Phonecall created	8	Phonecall created	8
295	Stage: Qualification → Leping	11	Phonecall created	11	Stage: Leping → Kohtumine/prese	11	Stage: Kohtumine/prese → Qualification	11	Stage: Qualification → Kohtumine/prese	11
391	Stage: Qualification → Kohtumine/prese	9	Expected revenue changed	20	Stage: Kohtumine/prese → Leping	20	Expected revenue changed	10	Expected revenue changed	10
592	Phonecall created	9	Stage: Qualification → Leping	9	Phonecall created	9	Phonecall created	9	Phonecall created	9

Table 16: 5-event matrix

5 Experimental Settings and Analysis

In this section, we describe briefly the experiments and the results acquired by classification training with different data mining techniques mentioned previously using predictive monitoring approach. We followed the following work-flow for predicting the cases in Order-to-contract process as early as possible.

- Step-1: Feature selection. The initial step is to recognize the variables for training the model for predicting the cases.
- Step-2: Identification of sequence length. Based on the variables selected in the Step-1, the sequence length is identified by comparison of the results.
- Step-3: Selection of the optimal model. Based on the best features defined in Step-1 and sequence length determined in step-2, different machine learning methods are applied to check which method performs better and check whether there is a model that shows better that the method used in Step-1 and Step-2 (Random Forest).

5.1 Feature selection

5.1.1 Which features to use

Based on the dataset features described in Section-4, we have large variables set that can be utilized for training. Although there are numerous features available, few features don't bring any importance to the model, so optimization of available features can deliver better quality in training and predicting the cases. This section focuses on the selection of best features that optimizes the design.

To verify the significance of the derived features, we analyzed the importance of the concluded features on positive and negative labels to identify the feature importance in the closed cases. Few features such as planned revenue, durations, entire events in the sequence play a significant role in the classification training when combined with the folded event data. After training every feature set with the Random Forest method, we retrieved the variable importance based on the trained model using R to check whether the features are important and which features are more important on the feature set. In Figure 9, Figure 10 and Figure 11 we interpret the importance of variables with duration in days.

Figure 9 showcases the importance of total events in the sequence on duration of the sequence in days. Positive and negative cases are mapped to identify the event lengths which are consuming more time and whether the positive and negative cases ended with different durations. Red colored lines represent the lost opportunities (negative cases), and the blue colored lines represent the won opportunities (positive cases). Sequence lengths from 1-12 can be identified as the considerable sequence lengths as the density is more in this area. We can quickly find the durations of positive and negative cases with different event periods from this plot, the event lengths with positive labels took more time to finish with maximum event lengths whereas in few cases such as with event lengths 7, 8, 12 negative cases took more time to finish. So from this correlation of event

lengths and duration, it is clear that event lengths play a significant role in the classification of positively and negatively ended cases.

Figure 9: Importance of entire events in the sequence with duration

In Figure 10, the importance of planned revenue and duration on won and lost opportunities. We can clearly observe the blue patch on the left side of the graph i.e. within the range of estimated revenue between 0-1000 more positive cases are located in this region. It is one of the unusual patterns that with lowest estimated revenue the cases ended with the positive outcome irrespective of duration. Another pattern can be seen in the planned revenue between 1000-3000 and length greater than 50 days which ended positive, here the interesting behavior is that the positive cases with higher planned revenues ended with long durations.

In contrast to the positive cases, negative cases can be observed in the region with the planned revenue between 1000-2000 and case duration less than 40 days. On the other side, there are few outliers with more estimated revenues in both the positive and negative cases which shows the sudden decrease and at the same time increase in durations.

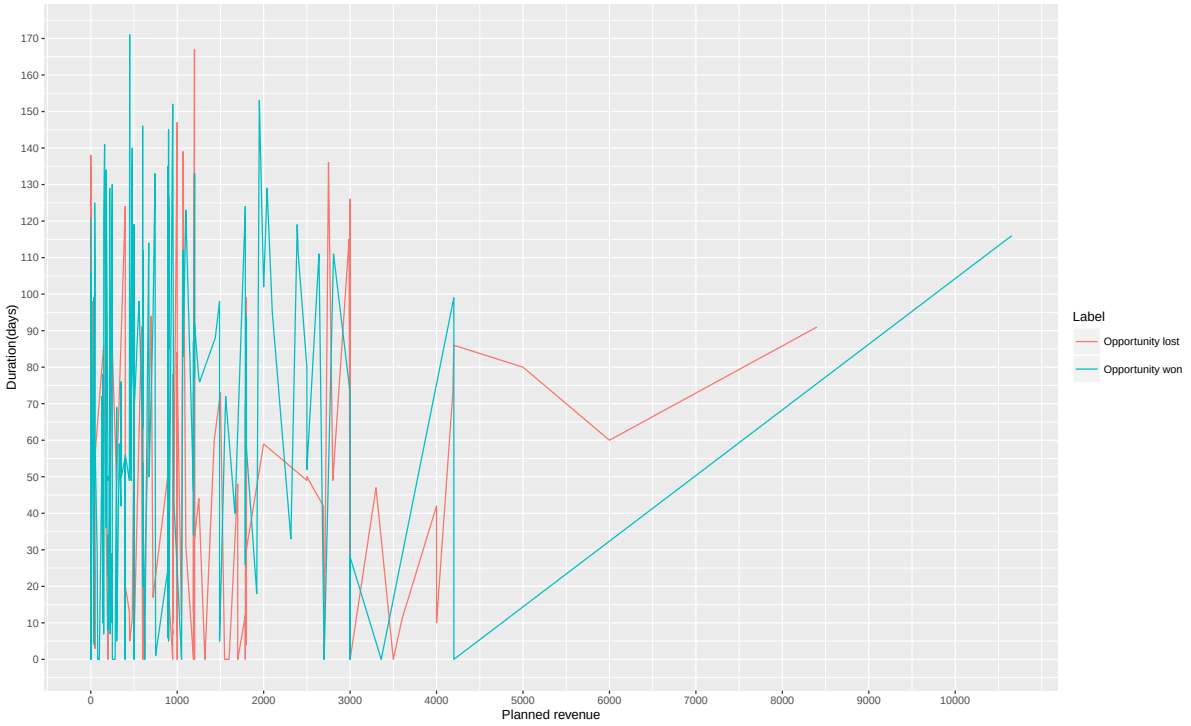

Figure 10: Importance of planned revenue with duration

Figure 11, represents the importances of the users feature with duration. User identification numbers of the sales representatives are displayed on the y-axis are the sales representatives in the company and duration displayed on the x-axis represents the case duration in days. In this graph, every process operators work can be noticed in the period of days, and a total number of positive and negative cases handled.

From Figure 11 it can be concluded that process operators (5, 8, 14, 15 and 20) handled a maximum of active cases. One interesting pattern that can also be found in Figure 11 is that the users 12, 16 and 21 ended up with more negative cases than positive cases and at the same they finished the negative cases faster and positive cases lately where one can identify a deviant behavior in the operation of these particular users. Another interesting pattern is that the process operators 25 and 26 ended up with only positive cases.

In overall positive cases took more duration compared to the adverse ones on process user. This characteristic of differentiating positive and negative scenario by stakeholder signifies the importance of the process operators in the success and failure of the opportunity in sales.

Lets us consider the n-event matrices found in Table 12, Table 13, Table 14, Table 15 and Table 16 can be represented using

$$\{e_1, e_2, e_3, e_4 ... e_n\}$$

Internal data features that are recored continuously in sales process such as expected revenue of the company, duration of the case, start date, and end date which is described

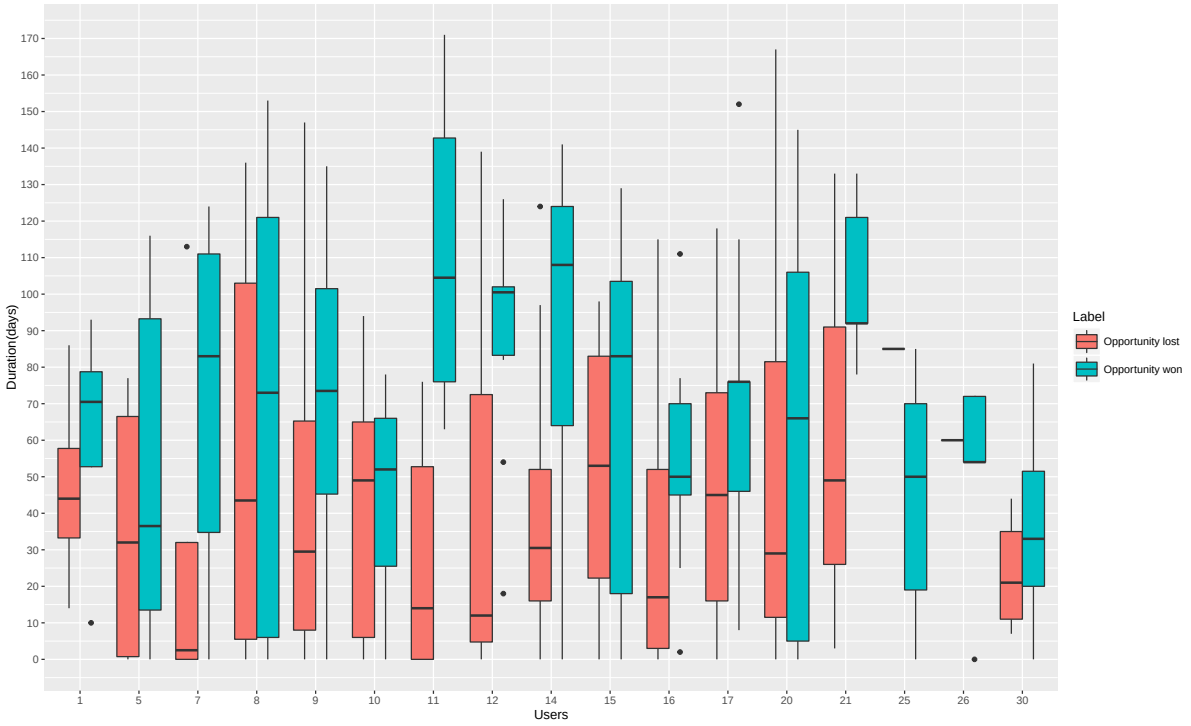

Figure 11: Importance of process users with duration

in Table 7 can be represented by

$$\{A_1, A_2, A_3, A_4, .., A_n\}$$

Frequency data shown in Table 10 can be represented by

$$\{F_1, F_2, F_3, ..F_n\}$$

External data described in Table 11 can be represented by

$$\{E_1, E_2, E_3, .., E_n\}$$

Figure 12 showcases the event length frequencies, where we can find the maximum occurrence of events lengths in the sequence. From the graph, it can be clearly seen that event lengths 2, 3, 4 and 5 are having maximal frequency of occurrence. In our study using predictive monitoring of identifying the cases as early as possible, we used the maximal frequent and initial sequence lengths for training as these event lengths has maximum cases after discarding the the finished labels inside the data to give early predictions.

In Figure 13, completed sequences are represented after discarding the finished events in the events such as 'Opportunity won', 'Opportunity lost', 'Dead' etc.,

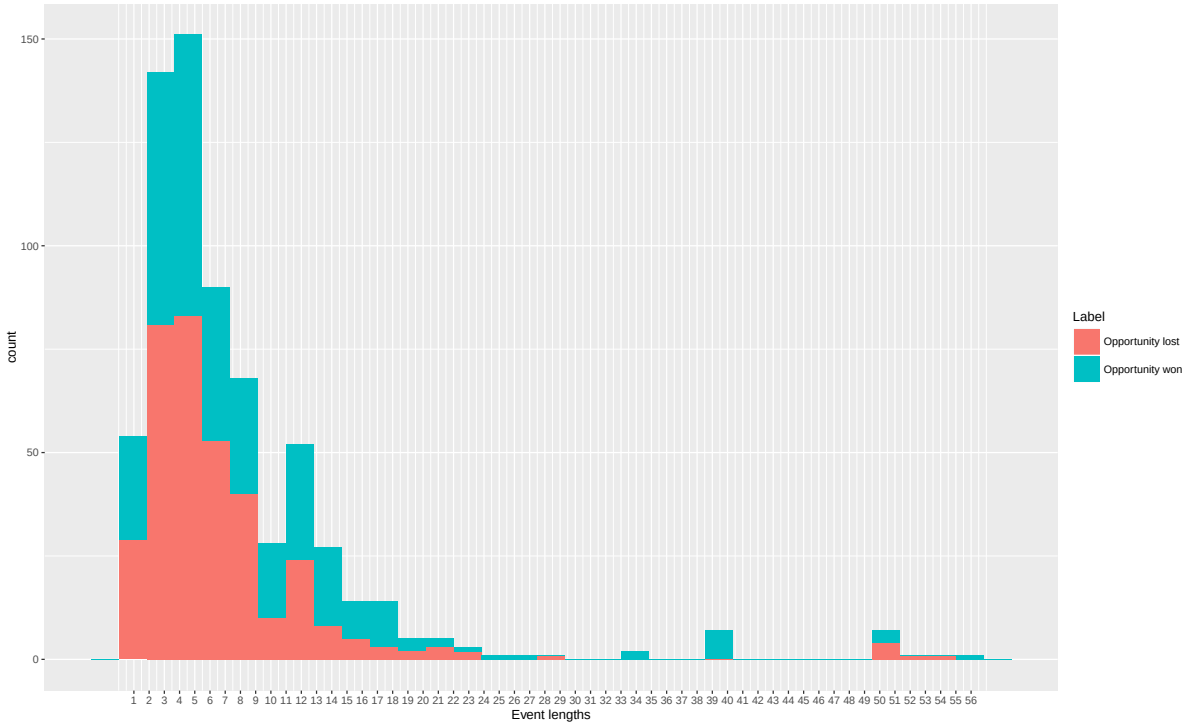

Figure 12: Overview of event lengths per instance and its frequency in event logs

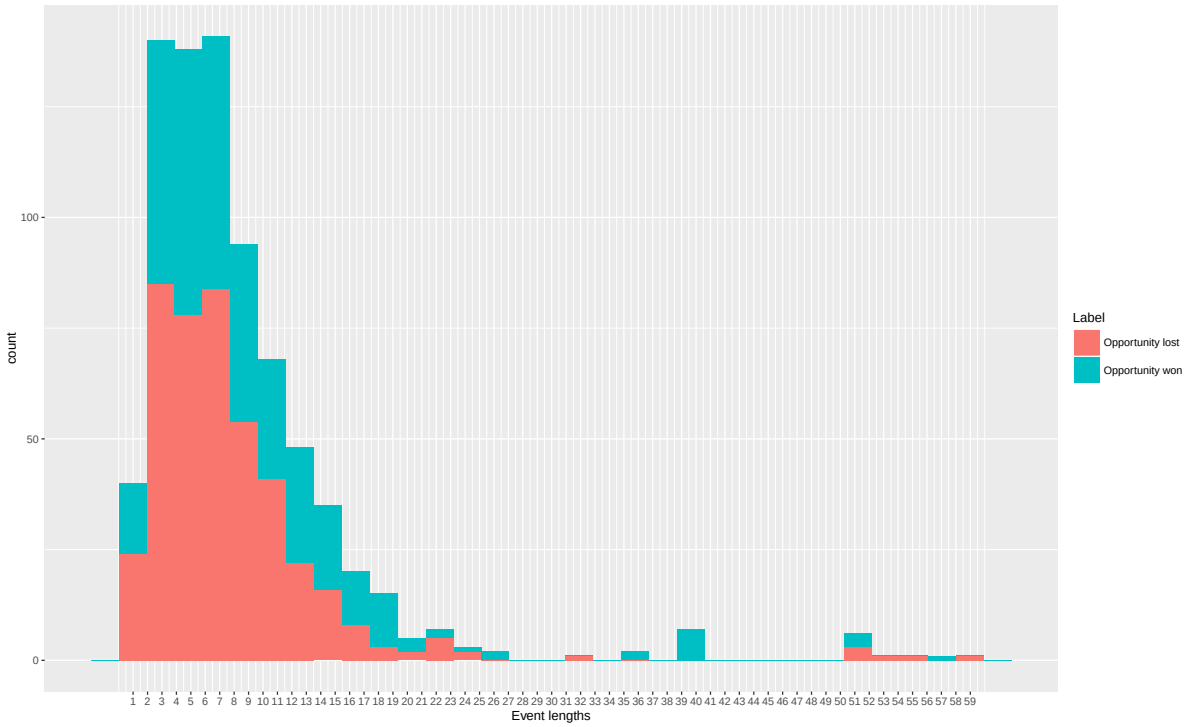

Figure 13: Overview of event lengths per instance and its frequency in event logs after discarding finished events

Based on the {1, 2, 3, 4, 5} event lengths the following are the possible combination

of features,

- Feature set-1: Only events,
$$\{e_1, e_2, e_3, e_4, e_5\}$$
- Feature set-2: Events, Internal data,
$$\{e_1 + A_1, e_2 + A_2, e_3 + A_3, e_4 + A_4, e_5 + A_5\}$$
- Feature set-3: Only External data,
$$\{E_1, E_2, E_3, E_4, E_5\}$$
- Feature set-4: Events , Internal data, External data,
$$\{e_1 + A_1 + E_1, e_2 + A_2 + E_2, e_3 + A_3 + E_3, e_4 + A_4 + E_4, e_5 + A_5 + E_5\}$$
- Feature set-5: Events, Internal data, Frequency data,
$$\{e_1 + A_1 + F_1, e_2 + A_2 + F_2, e_3 + A_3 + F_3, e_4 + A_4 + F_4, e_5 + A_5 + F_5\}$$
- Feature set-6: All the above features combined,
$$\{e_1+A_1+E_1+F_1, e_2+A_2+E_2+F_2, e_3+A_3+E_3+F_3, e_4+A_4+E_4+F_4, e_5+A_5+E_5+F_5\}$$

Assuming that the 5-event matrix is better regarding predictive monitoring approach as it is equivalent to the mean of the event lengths and captures the sequence. We conduct our experiments on this event fold and check for the features that are responsible for giving optimal solutions. All the data features are combined with 5-fold event matrix to select the best features and check whether only external data without predictive monitoring problem gives best results or not. After combining the features sets with the 5-fold event matrix as mentioned in Section 4.3, we classify the finished and unfinished cases separately using the label(won/lost) in event logs with a binary classifier. Then we apply the predictive monitoring approach on the 5-event matrix with different data features

We trained the model using the 5-event matrix with Feature set-1, Feature set-2, Feature set-3, Feature set-4, Feature set-5 and Feature set-6 with a data split of 50% of data for training and the remaining for testing because of smaller dimensions of data. We then compared the results based on F-score and Accuracy to chose the optimal feature set.

Feature set-1:
$$\{e_5\}$$
Feature set-2:
$$\{e_5 + A_5\}$$
Feature set-3:
$$\{E_1, E_2, E_3, E_4, E_5\}$$
Feature set-4:
$$\{e_5 + A_5 + E_5\}$$
Feature set-5:
$$\{e_5 + A_5 + F_5\}$$
Feature set-6:
$$\{e_5 + A_5 + E_5 + F_5\}$$

5.1.2 External data vs. predictive monitoring

We run our experiments on all the six different feature vectors on 5-event matrix to check which model gives the optimal solution. In this subsection, we also demonstrate why we consider predictive monitoring(Feature set-1, Feature set-2, Feature set-4, Feature set-5 and Feature set-6) data structure other than external data(Feature set-3). Figure 11 and Figure 12 shows the results of all the feature sets with 5-event matrix with respective F-score and Accuracy. External data features have no relation with the event logs; we trained model also to test and identify if only external data considered would give better results. Better scores can be seen clearly in the cases with Feature set's associated with the predictive monitoring approach, i.e., including the event log data as a combination without considering any external data.

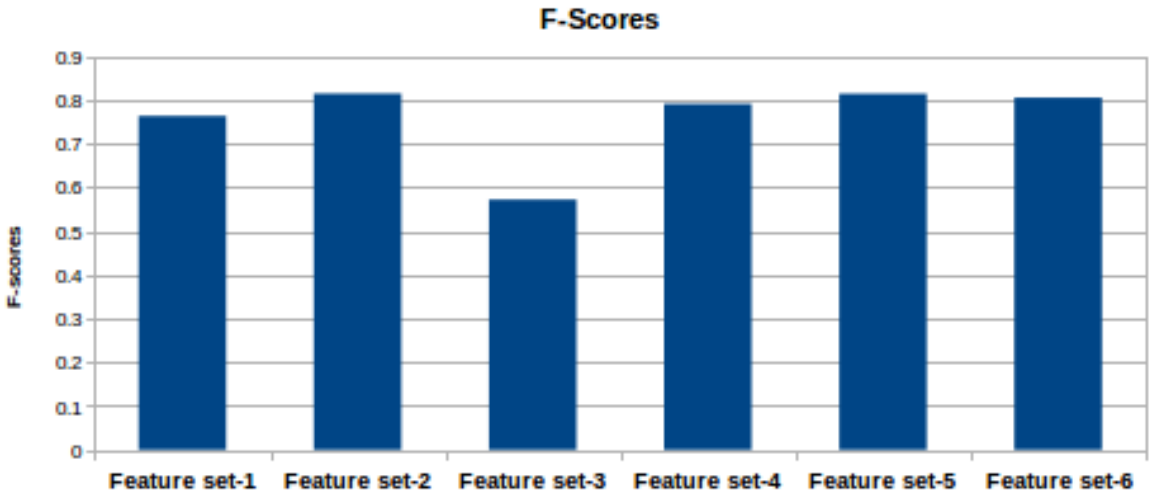

Figure 14: Comparison of F-scores of different feature sets

Figure 14 clearly represents that the Feature set-2 and Feature set-5 has the highest F-score where as Feature set-3 has the least F-score. Feature set-1, Feature set-2, Feature set-4, Feature set-5 and Feature set-6 are event wrapped and they are based on predictive monitoring problem for predicting the cases as early as possible. Feature set-3 has only the external data features. In Figure 15, the accuracies results are displayed for the various available Feature sets. Feature set-5 has the maximum accuracy whereas the feature set-3 has the least accuracy levels.

Feature sets with predictive monitoring method shows the steady F-scores which are more than 70% whereas in contrast to these external data variables (Feature set-3) shows the weakest which is near 55% of F-score. Figure 12 clearly indicates that the similar scores observed in Figure 11, i.e. Feature set-2 and Feature set-5 has the highest Accuracy whereas Feature set-3 has the least accuracy. Accuracies of available feature sets displayed in Figure 12 conveys us that predictive monitoring approach has a margin of 60% of Accuracy in contrast to the external data features which is considerably low contributing to 45% of Accuracy.

One can identify the importance of these features by merging the feature sets with different dimensions as there might be a chance of change in the results with identified feature set. More understanding and confirmation of features can also be established from the study of these features combined with different event folds which is explained in Section 5.2

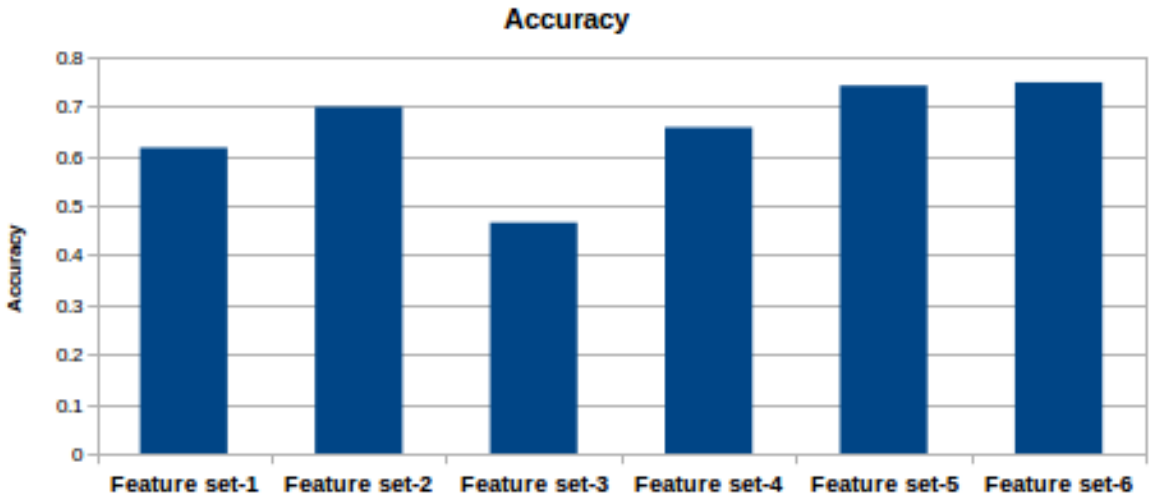

Figure 15: Comparison of accuracies of different feature sets

Based on the F-scores and Accuracy results of different feature sets, the model with external data in Feature set-3 parameters shows the weaker results, and the results with models based on predictive monitoring problem show the best results. Feature set-2 and Feature set-5 which are based internal data features shows the best performance without any external data. Hence from the results displayed in Figure 14 and Figure 15, we conclude that predictive monitoring approach with Feature set-5 is considered as the most desirable features.

5.2 Identification of sequence length

In the previous section we identified the features based on the 5-fold event matrix, but in general, we don't know which n-fold matrix gives us the best results. In this section to identify the best sequence length suitable for the feature sets, we run our experiments with the event types the 1-event matrix, 2-event matrix, 3-event matrix, 4-event matrix, and 5-event matrix. After combining the Feature set-1, Feature set-2, Feature set-3, Feature set-4, Feature set-5 and Feature set-6, we train the model and compare the results obtained for specified datasets and decide which data model gives us the best fit. In this section, we determine which type (n-event matrix) should be considered rationally.

For training purpose, we used "Random Forest" method of classification to predict the outcome of the case. To train the model we used 10-fold cross-validation technique with 50% split ratio for training and testing datasets as the sample size is considerably low especially with the bigger event lengths. In Figure 16, Figure 17, Figure 18 and Fig-

ure 19 we present model results based on F-score, Accuracy, Precision and Recall metrics.

In Figure 16 we can see the training results based on F-scores for all the feature sets and event folds can be observed. Results displayed in descending order from left to right with highest scores on the left and the least scores on the right using the average score of the feature sets for every event fold. Results clearly indicate that the 3-event matrix, 4-event matrix, 5-event matrix showed us the best-performing metrics across all the Feature sets. Best results can be seen in 3-event matrix with Feature set-5. Also similar results can be seen in Feature set-2 and Feature set-1. From this, we conclude that features with event data matrix and internal data features(Revenue, Durations, Frequency encoded data) shows us the best performing combinations.

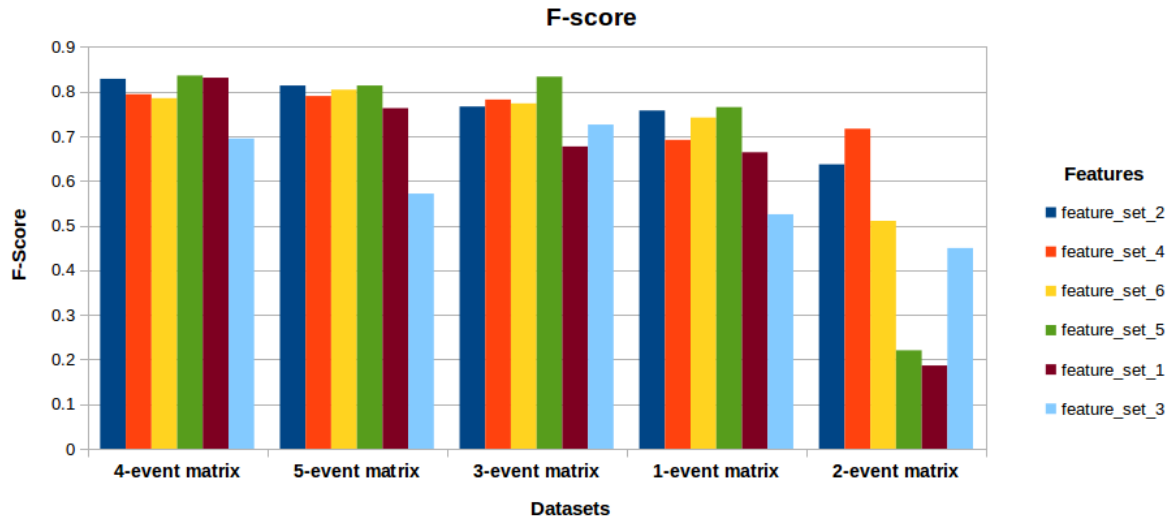

Figure 16: Comparison of F-score metric results with n-event matrices

In Figure 17, the best-performing combination of datasets according to accuracy metric can be seen in 3-event matrix and Feature set-5. Similar to F-scores, Accuracies of all the combinations indicate that the best combinations can be found in 3-event matrix, 4-event matrix and 5-event matrix. As discussed in Section 5.1, predictive monitoring approach gives best results in all the event wraps except 2-event matrix, but the margin of Accuracy and F-score is below 50%.

In Figure 18 and Figure 19 the possible combinations of datasets shows the higher Precision and lower Recall because of the model estimates results for negatively classified cases more. Although the 3-event matrix, 4-event matrix and 5-event matrix are the best performing according to Precision and Recall.

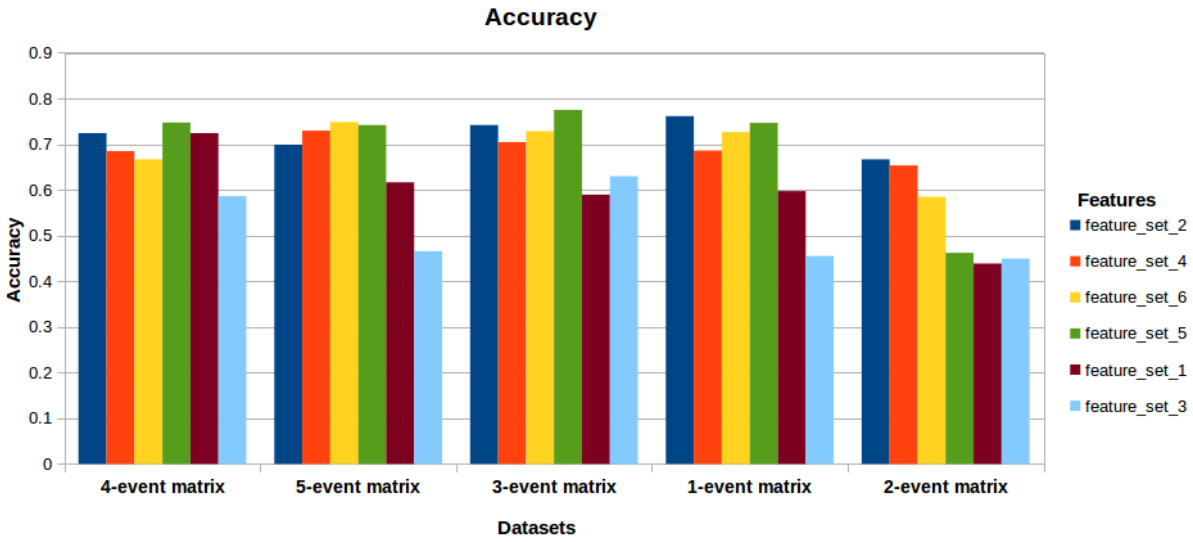

Figure 17: Comparison of Accuracy metric results with n-event matrices

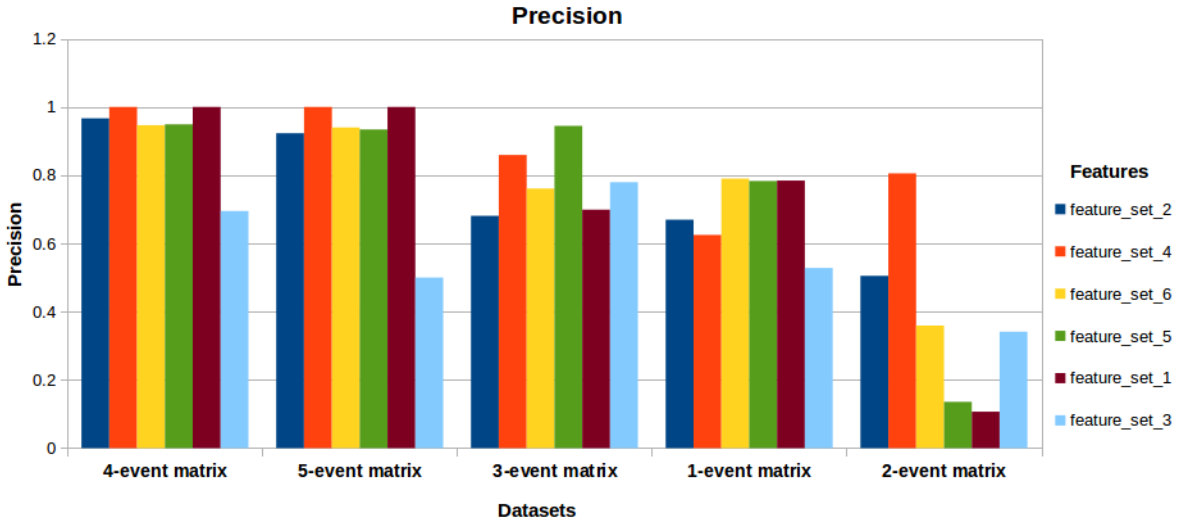

Figure 18: Comparison of Precision metric results n-event matrices

Based on the results using different metrics, we concluded to use the 3-event matrix and Feature set-5 in our experiments as the metrics are higher in this case. Other advantage of using the 3-event matrix is that it has more cases in the dataset compared to 4-event matrix and 5-event matrix.

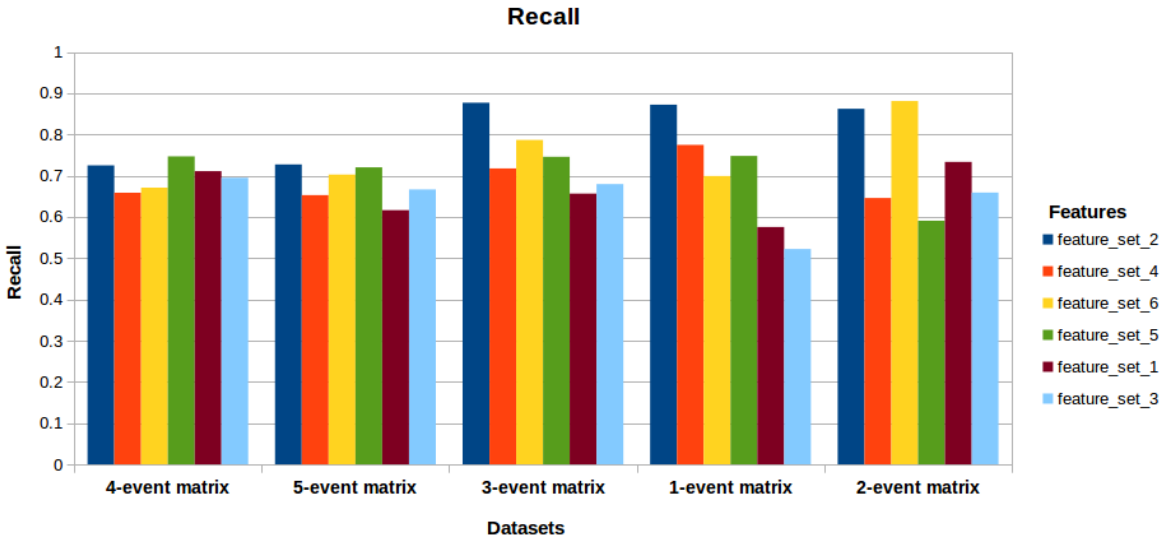

Figure 19: Comparison of Recall metric results with n-event matrices

5.3 Selection of the optimal model

In this section, we select the best model that fits the predictive monitoring problem for predicting the cases as early as possible. We experimented our data with different machine learning methods available in R, but only a few methods are compatible with our data structure. The following methods are consistent with our datasets. In this section, we used 3-event matrix with 'Feature set-5' to find the best performing method for our work.

- Random Forest (RF)
- eXtreme Gradient Boosting (EGB)
- Single C5.0 Ruleset (C5.0 Rules)
- Naive Bayes (NB)
- Penalized Discriminant Analysis (PDA)
- Principle least squares (PLS)
- Support vector machines (SVM)

Due to the lesser number of cases available in the dataset we have chosen the 50% for training the cases and the remaining 50% of the cases for testing purpose. We used 10-fold cross-validation to train our model. We measured and selected the best methods based on Accuracy, F-scores, Precision and Recall of these methods to conclude the best performing method. Figure 20 presents the F-scores of mentioned methods, among all the methods available, Random Forest has the highest F-score.

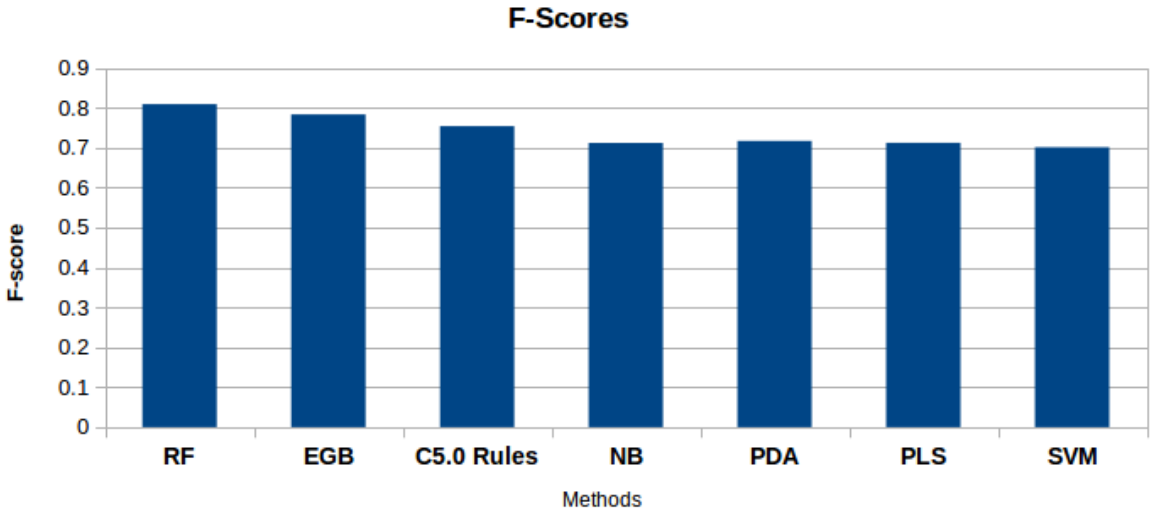

Figure 20: Comparision of F-score of different models

Figure 21 represents the Accuracies obtained after training the model with the methods mentioned earlier. Similar to the F-scores of these methods, accuracies shows the highest in terms of Random Forest method for classification of positive and negative cases.

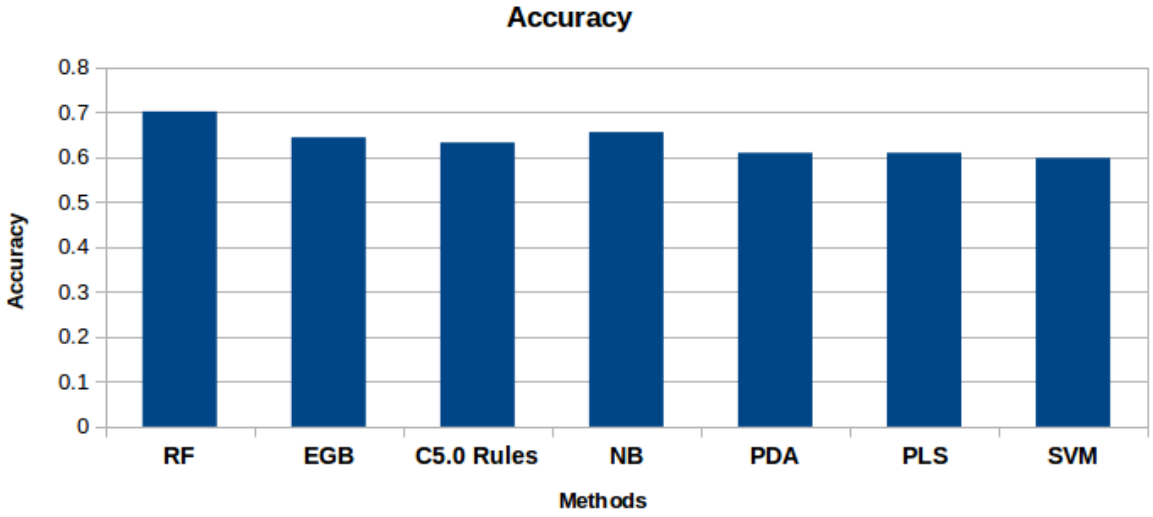

Figure 21: Comparision of Accuracy of different models

Figure 22 and Figure 23 represent the Precision and Recall results of every method in the list respectively. Percentage of positive case predictions were correctly classified in RF and EGB according to the precision. More positive cases are predicted using NB and RF as the we see that Recall is higher both of these cases. According to precision and recall, the methods (RF, EGB, NB) predicted the positive cases well.

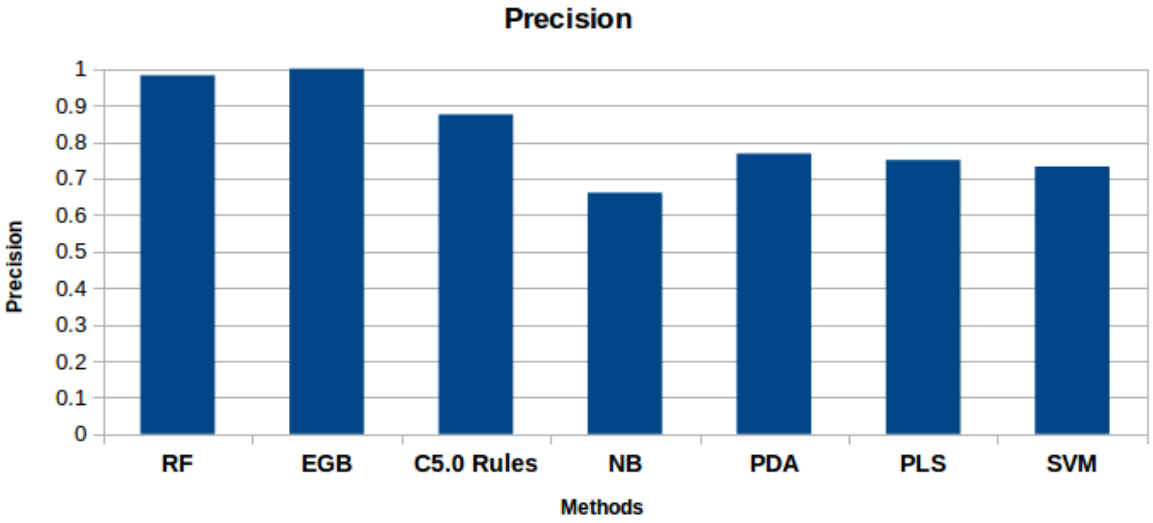

Figure 22: Comparision of Precision of different models

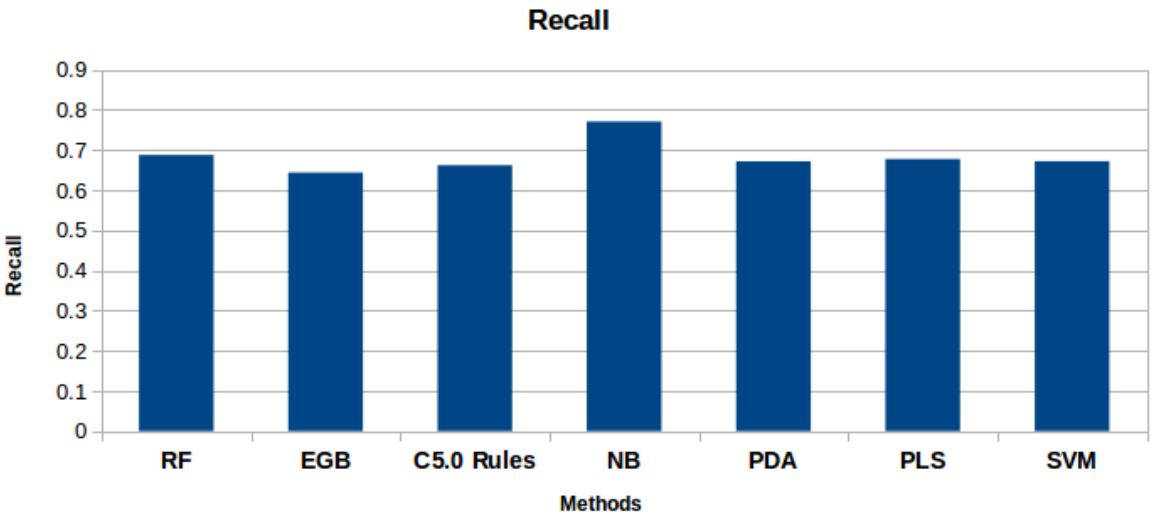

Figure 23: Comparision of Recall of different models

Although the model training results show that the RF method is efficient in giving better F-score and Accuracy, other methods such as EGB and C5.0 Rules also provide better results, and they can be considered for understanding the data. As we were unable to design decision trees for more insight of the chosen data, we used C5.0 Rules for constructing the decision tree.

Based on then training results using different parameters(F-score, Accuracy, Precision and Recall) clearly indicate that the 'Random Forest' method is the best performing method among the experimented methods. So we decided to use Random Forest for our experiments and practical usage for predicting the cases as early as possible.

5.4 Threats to validity

This research work is based on a relatively small sample of 1000 CRM cases from November 2015 to April 2016. A more extensive study with bigger samples of cases is desirable for a better identifying of how predictive monitoring is helpful in predicting the CRM cases as early as possible. The results of this research work could be affected by the unbalanced dataset, but there are not many solutions to this threat other than using a larger number of samples.

Another potential threat is that the external data(Feature set-3) when combined with the n-event matrix together and separately are giving the least scores which can be seen in Figure 16, Figure 17, Figure 18 and Figure 19. Maybe with more sample size they might give better results when combined with event folds. For a better understanding of external data usable with the event wraps a more extensive study could be helpful in how the external data added to the event folds would improve the results and usable as a feature set.

This study did not cover all possible data mining techniques, as only the available and the most used methods in the selected R toolset. This study included different methods which were discussed in Section 5.3, although Random Forest gives best results in different domains. For example in some cases Naive Bayes gives the best results but in our case it gave the worst results. Therefore, dataset and selected features could deliver better results with some other methods. Due to limited time, we could not experiment all the methods available via different tools and techniques therefore if those methods are included could improve the results.

Finally, due to the previously mentioned potential threats, one should not depend only on computed predictive monitoring probabilities using the proposed procedure in the previous section when making business decisions because a greater number of samples are needed to confirm the applicability of the model for practical use. However, it should be safe to use the probability as an indicator of possible alerts for identifying the risky opportunities beforehand.

6 Conclusion

Predictive monitoring is a necessary approach for signifying the business decisions and profitability and is a well-researched area. In this research, we investigated the combination of external data with event logs. The empirical results showed that external data when combined with the event data, had the negative effect on the model performance.

However, internal data, which was collected during the sales process together with the event data gave us the best results. Results also conveyed that using three first events in the event logs and internal data which has high correlation produced the optimal model performance based on the Accuracy and F-score. Furthermore, Random Forest was the most suitable method for our datasets based on the model comparison results. Finally, we proposed the ranking of opportunities to the sales funnel based on the probabilities obtained from the model for unfinished cases.

Overall we developed a model for ranking the opportunities in the sales funnel and analyzed the optimal features and model for the best solution. Although the practical implication of results needs to be verified. Few challenges should also be addressed such as the limitation of data in event logs with the lesser number of cases. To check for the optimal performance of the model, one should use bigger datasets with balanced positive and negative cases to verify the model. Also, the negativity of the external data features performance was unknown; it is recommended to investigate the external data features with the event logs that shown the weakest performance in this study.

Acknowledgements:

First, I would wish to express my gratitude to my supervisor Dr. Peep Küngas for the endless support of my Master's thesis study, for his patience, motivation, and tremendous knowledge. His guidance encouraged me in all the time of research and writing of this thesis.

Besides I would also thank IT Akadeemia for supporting my studies by sponsoring scholarship for my Master's study.

Last but not least, I like to thank my family: my parents and friends for supporting me throughout my life.

References

[1] Leontjeva, A., Conforti, R., Di Francescomarino, C., Dumas, M., & Maggi, F. M. (2015). Complex Symbolic Sequence Encodings for Predictive Monitoring of Business Processes. In *Business Process Management* (pp. 297-313). Springer International Publishing.

[2] Di Francescomarino, C., Dumas, M., Maggi, F. M., & Teinemaa, I. (2015). Clustering-Based Predictive Process Monitoring. *arXiv preprint arXiv:1506.01428.*

[3] Verenich, I., Dumas, M., La Rosa, M., Maggi, F. M., & Di Francescomarino, C. (2015). Complex symbolic sequence clustering and multiple classifiers for predictive process monitoring.

[4] Van der Aalst, W. M., Schonenberg, M. H., & Song, M. (2011). Time prediction based on process mining. In *Information Systems*, 36(2), 450-475.Chicago

[5] Sebu, M. L., & Ciocarlie, H. (2015, May). Business activity monitoring solution to detect deviations in business process execution. In *Applied Computational Intelligence and Informatics (SACI), 2015 IEEE 10th Jubilee International Symposium* (pp. 437-442). IEEE.

[6] Gallagher, C., Madden, M. G., & D'Arcy, B. (2015, December). A Bayesian Classification Approach to Improving Performance for a Real-World Sales Forecasting Application. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)* (pp. 475-480). IEEE.

[7] Schonenberg, H., Weber, B., Van Dongen, B., & Van der Aalst, W. (2008). Supporting flexible processes through recommendations based on history. In *Business process management* (pp. 51-66). Springer Berlin Heidelberg.

[8] Zeng, S., Melville, P., Lang, C. A., Boier-Martin, I., & Murphy, C. (2008, August). Using predictive analysis to improve invoice-to-cash collection. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1043-1050). ACM.

[9] Gróger, C., Schwarz, H., & Mitschang, B. (2014, May). Prescriptive analytics for recommendation-based business process optimization. In *In Business Information Systems* (pp. 25-37). Springer International Publishing.

[10] Xing, Z., Pei, J., & Keogh, E. (2010). A brief survey on sequence classification. In *ACM SIGKDD Explorations Newsletter*, 12(1), 40-48.

[11] Xing, Z., Pei, J., Dong, G., & Philip, S. Y. (2008, April). Mining Sequence Classifiers for Early Prediction. SDM (pp. 644-655).

[12] Nguyen, H., Dumas, M., La Rosa, M., Maggi, F. M., & Suriadi, S. (2014, October). Mining business process deviance: a quest for accuracy. In *On the Move to Meaningful Internet Systems: OTM 2014 Conferences* (pp. 436-445). Springer Berlin Heidelberg.

[13] Dumas, M., & Maggi, F. M. (2015). Enabling process innovation via deviance mining and predictive monitoring. In *BPM-Driving Innovation in a Digital World* (pp. 145-154). Springer International Publishing.

[14] Liang, Y., Zhang, Y., Xiong, H., & Sahoo, R. (2007, October). Failure prediction in ibm bluegene/l event logs. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference* on (pp. 583-588). IEEE.

[15] Pika, A., van der Aalst, W. M., Fidge, C. J., ter Hofstede, A. H., & Wynn, M. T. (2012, September). Predicting deadline transgressions using event logs. In *Business Process Management Workshops* (pp. 211-216). Springer Berlin Heidelberg.

[16] Pika, A., Van Der Aalst, W. M., Fidge, C. J., Ter Hofstede, A. H., & Wynn, M. T. (2013, June). Profiling event logs to configure risk indicators for process delays. In *Advanced Information Systems Engineering* (pp. 465-481). Springer Berlin Heidelberg.

[17] Song, J., Luo, T., & Chen, S. (2008, April). Behavior pattern mining: Apply process mining technology to common event logs of information systems. In *Networking, Sensing and Control, 2008. ICNSC 2008. IEEE International Conference* on (pp. 1800-1805). IEEE.

[18] van der Aalst, W. M., Low, W. Z., Wynn, M. T., & ter Hofstede, A. H. (2015, May). Change your history: Learning from event logs to improve processes. In *Computer Supported Cooperative Work in Design (CSCWD), 2015 IEEE 19th International Conference* (pp. 7-12). IEEE.

[19] Bose, R. P., & van der Aalst, W. M. (2013, April). Discovering signature patterns from event logs. In *Computational Intelligence and Data Mining (CIDM)*, 2013 IEEE Symposium on (pp. 111-118). IEEE.

[20] Gehrke, N., & Werner, M. (2013). Process mining. In *Das Wirtschaftswachstum*, 42 (7), 934-943.

[21] van der Aalst, W. M., Reijers, H. A., Weijters, A. J., van Dongen, B. F., De Medeiros, A. A., Song, M., & Verbeek, H. M. W. (2007). Business process mining: An industrial application. In *Information Systems*, 32(5), 713-732.

[22] Chen, I. J., & Popovich, K. (2003). Understanding customer relationship management (CRM) People, process and technology. In *Business process management journal*, 9(5), 672-688.

[23] Reinartz, W., Krafft, M., & Hoyer, W. D. (2004). The customer relationship management process: Its measurement and impact on performance. In *Journal of marketing research*, 41(3), 293-305.

[24] Farooqi, M., & Raza, K. (2012). A Comprehensive Study of CRM through Data Mining Techniques. arXiv preprint arXiv:1205.1126.

[25] 10X more productive series. Understand your pipeline and track sales effectively - Base CRM Blog, In *http://downloads.getbase.com/base-crm-reading-material/understanding-your-pipeline-and-sales-tracking.pdf*

[26] Original version by Odoo, Odoo Customer Relationship Management(Release8), In *URL https://www.odoo.com/*, November 03, 2014.

[27] Affinity Service Definition: OpenERP Odoo Customer Relationship Management, In *URL http:www.affinity-digital.com*

[28] Best Practices for Sales Managers: Salesforce.com. In *URL http://www.salesforce.com/assets/pdf/misc/BP_SalesManagers.pdf*

[29] Janssenswillen, G., Swennen, M., Depaire, B., Jans, M., & Vanhoof, K. (2015). Enabling Event-data Analysis in R: Demonstration.

[30] Kosorus, H., Hönigl, J., & Küng, J. (2011, August). Using r, weka and rapidminer in time series analysis of sensor data for structural health monitoring. In *Database and Expert Systems Applications (DEXA), 2011 22nd International Workshop on* (pp. 306-310). IEEE.

[31] Kumar, P., Ozisikyilmaz, B., Liao, W. K., Memik, G., & Choudhary, A. (2011, May). High performance data mining using R on heterogeneous platforms. In *Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011 IEEE International Symposium* (pp. 1720-1729). IEEE.

[32] Kehrer, J., Boubela, R. N., Filzmoser, P., & Piringer, H. (2012, October). A generic model for the integration of interactive visualization and statistical computing using R. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference* (pp. 233-234). IEEE.

[33] Zakirov, D., Bondarev, A., & Momtselidze, N. (2015, September). A comparison of data mining techniques in evaluating retail credit scoring using R programming. In *2015 Twelve International Conference on Electronics Computer and Computation* (ICECCO) (pp. 1-4). IEEE.

[34] Chen, T., & He, T. (2015). xgboost: eXtreme Gradient Boosting. R package version 0.4-2.

[35] Kuhn, M., Weston, S., Coulter, N., & Quinlan, R. (2014). C50: C5. 0 decision trees and rule-based models. In *R package version 0.1. 0-21*, URL http://CRAN. R-project. org/package C, 50.

[36] Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized discriminant analysis. In *The Annals of Statistics*, 73-102.

[37] Tobias, R. D. (1995, April). An introduction to partial least squares regression. In *Proc. Ann. SAS Users Group Int. Conf.*, 20th, Orlando, FL (pp. 2-5).

[38] Meyer, D., Leisch, F., & Hornik, K. (2003). The support vector machine under test. In *Neurocomputing*, 55(1), 169-186.

[39] Van der Aalst, W. M., van Dongen, B. F., GÃijnther, C. W., Rozinat, A., Verbeek, E., & Weijters, T. (2009). ProM: The Process Mining Toolkit. In *BPM (Demos)*, 489, 31.

[40] Jaroenphol, E., Porouhan, P., & Premchaiswadi, W. (2015, November). Analysis of the patients' treatment process in a hospital in Thailand using fuzzy mining algorithms. In *ICT and Knowledge Engineering (ICT & Knowledge Engineering 2015), 2015 13th International Conference* (pp. 131-136). IEEE.

[41] Jaisook, P., & Premchaiswadi, W. (2015, November). Time performance analysis of medical treatment processes by using disco. In *ICT and Knowledge Engineering (ICT & Knowledge Engineering 2015), 2015 13th* International Conference (pp. 110-115). IEEE.

Appendix

A. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I **Madhu Tipirishetty** (DOB: 14-08-1989),
(*author's name*)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Predictive process monitoring for Lead-to-Contract process optimization,
(*title of thesis*)
supervised by **Peep Küngas**,
(*supervisor's name*)

2 I am aware of the fact that the author retains these rights.

3 I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tallinn, **May 19, 2016**