REVA
UNIVERSITY
Bengaluru, India

A Project Report on

Predicting Dispute Status using Machine Learning Approach

Submitted in Partial Fulfilment for Award of Degree of

Master of Business Administration
In Business Analytics

Submitted By
Madhukeshwar. R K
R19MBA57

Under the Guidance of
Mr. Dipanjan Deb
VP, Wells Fargo

REVA Academy for Corporate Excellence - RACE
REVA University
Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064
race.reva.edu.in

March, 2021

REVA
UNIVERSITY
Bengaluru, India

Candidate’s Declaration

I, **Madhukeshwar R K** hereby declare that I have completed the project work towards the first year of Master of Business Administration in Business Analytics at, REVA University on the topic entitled **Predicting Dispute Status using Machine Learning Approach** under the supervision of **Mr. Dipanjan Deb**. This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year 2021.

Madhukeshw RK

Place: Bengaluru
Date: 6th March 2021

Name of the Student: Madhukeshwar R K
Signature of Student

REVA
UNIVERSITY
Bengaluru, India

Certificate

This is to Certify that the Project work entitled **Predicting Dispute Status using Machine Learning Approach** carried out by **Madhukeshwar R.K** with SRN R19MBA57, is a bonafide student of REVA University, is submitting the first year project report in fulfilment for the award of **Master of Business Administration** in Business Analytics during the academic year 2021. The Project report has been tested for plagiarism and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of PROJECT work prescribed for the said Degree.

Dipanjan Deb

Signature of the Guide	Signature of the Director
Name of the Guide **Mr. Dipanjan Deb**	Name of the Director **Dr. Shinu Abhi**
Guide	Director

External Viva

Names of the Examiners

1. Indrajit Kar, Head of AI-Chief Architect & Data Scientists, Siemens
2. Pradeepta Mishra, Associate Principal & Head of AI, LTI -Larsen and Toubro Infotech.

Place: Bengaluru

Date: 6th March 2021

REVA
UNIVERSITY
Bengaluru, India

Acknowledgement

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

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

I would like to acknowledge the support provided by Hon'ble Chancellor, **Dr. P Shayma Raju**, Vice Chancellor, **Dr. K. Mallikharjuna Babu,** and Registrar, **Dr. M. Dhanamjaya**. It is sincere thanks to all members of program office of RACE who were supportive in all requirements from the program office.

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

Place: Bengaluru
Date: 6th March 2021

REVA
UNIVERSITY
Bengaluru, India

Similarity Index Report

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

Software Used: Turnitin
Date of Report Generation: 3rd March 2021
Similarity Index in %: 14%
Total word count: 4585
Name of the Guide: Mr. Dipanjan Deb

Madhukeshwar RK

Place: Bengaluru
Date: 3rd March 2021
Verified by: Andrea Brian C

Name of the Student: Madhukeshwar R K
Signature of Student

Signature
Dr. Shinu Abhi,
Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	CRISP-DM	Cross-Industry Process for Data Mining
2	EDA	Exploratory Data Analysis
3	AR	Accounts Receivable
4	LR	Logistic Regression
5	CART	Decision Tree Classifier
6	NB	Naïve Bayes
7	SVM	Support Vector Machine
8	ROC	Receiver Operating Characteristic

List of Figures

List of Tables

Abstract

The scope of this project is to identify well in advance, the customers disputed invoices get rejected or approved, so that dispute management team can initiate conversions with these customers and try to address their concern even before the disputed invoice gets rejected. The data is mostly in a structured format capturing the lifecycles of disputes. This would form the primary dataset of our study.

The business impact of this project will lead to reducing the number of dispute rejection and to get the invoice amount paid by the customer well within the time period. This study also led to identifying the latent patterns regarding raising the disputes.

Keywords: Order to Payment Cycles, Invoice Disputes, Customer Analysis, Predictive Modeling.

Contents

Chapter 1: Introduction

Order to Cash business process involves account receivable collections after an invoice is issued to the customer (Cheong et al., 2018). Invoices are used where services and product are provided and they usually contain the rendered charges(Fernandez & Yuan, 2010). Typical payment terms provided would be of 30, 45 and 60 days to customer to make full payment of the invoiced amount. However, in business certain customers do not make invoiced amount on time and an intervention actions to remind their customer is required, this involves cost, money and time even lead to poor customer satisfaction (Cheong et al., 2018).

Standard 30-day payment term is provided to customers to make a payment of invoiced amount and 45 days of allowance is permitted before the intervention action starts. Soft intervention reminders of emails, messages are sent after 45 days and hard intervention of demand letter post 60 days. After 180 days, payment amount will be deemed as bad debt and no future orders will be accepted from such bad customers (Cheong et al., 2018).

During the invoice process a dispute can be raised by the customer for the invoice which could be due to product mismatch or the exceptions of the delivery of the product is not meet. Certain customers raise false dispute in order to gain additional time to make the invoiced amount. Unpleasant dispute over invoice amount could take place, which could even require legal resolution.

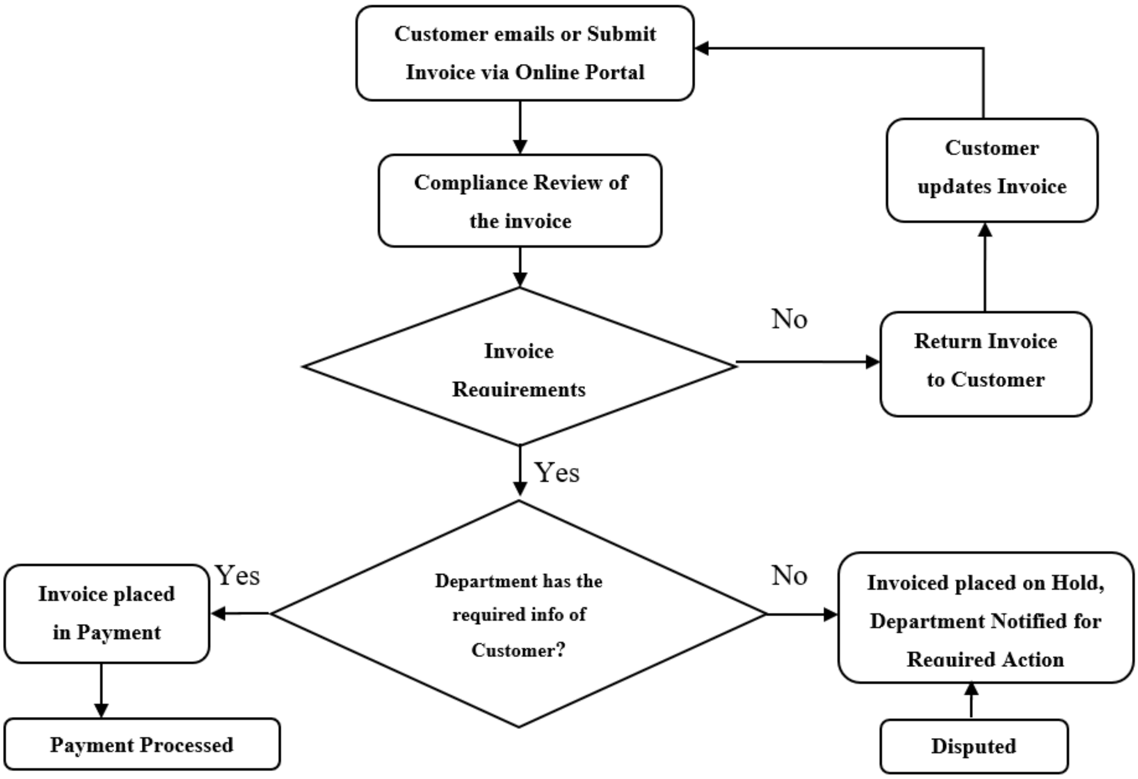

Customer emails or Submit Invoice via Online Portal
Compliance Review of the invoice
Invoice Requirements
No
Return Invoice to Customer
Customer updates Invoice
Yes
Department has the required info of Customer?
Yes
Invoice placed in Payment
Payment Processed
No
Invoiced placed on Hold, Department Notified for Required Action
Disputed

Figure 1.1- Invoice Process Cycle

Chapter 2: Literature Review

There are many conceptual and empirical studies available in the extant literature on improvising the order to cash cycle process, especially on Accounts Receivables. One of the studies by Fernandez and Yuan suggests to analyze the pattern for invoice processing. The pattern describes events such as the creation and validation of an invoice, followed by the payment process. This pattern is composed of two simpler patterns that describe the creation and payment of the invoice. The component patterns have value of their own and can be used independently(Fernandez & Yuan, 2010)**.**

"It is commonly agreed that AR (account receivable) is most valuable asset of any business firm. It can be source of financial difficulties for firm when they are not efficiently managed and underperforming. So, it is important to identify data pattern in AR and get meaningful insight from AR data". This paper demonstrates how supervised machine learning can help to build model to predict payment outcome of invoices which are yet not paid (Open) based on historical data(Shah, 2019).

"One of the main costs associated with Accounts receivable (AR) collection is related to the intervention actions taken to remind customers to pay their outstanding invoices. Apart from the cost, intervention actions may lead poor customer satisfaction, which is undesirable in a competitive industry"(Cheong et al., 2018)

"The account receivable is one of the main challenges in the business operation. With poor management of invoices to cash collection process, the overdue invoice may pile up, and the increasing amount of unpaid invoice may lead to cash flow problems. In this thesis, I addressed the proactive approach to improve account receivable management using predictive modeling"(Hu, 2009)

"We are interested in improving AR collection through machine learning for three reasons.
First of all, AR collection can easily be a source of financial difficulty of firms, if not well managed. It is, therefore, of great interests to manage it more effectively. Also, most of the AR collection actions nowadays are still manual, generic and expensive. For instance, it seldom takes into account customer specifies, neither has any prioritizing strategies.

Last and most importantly, commercial firms now are accumulating large amount of data about their customers, which makes the large-scale data-driven AR collection possible.”(Peiguang, 2015)

A study by Tater and others, have developed a classification model to identify the delayed invoices as a supervised classification task(Tater et al., 2018).

“Experience across multiple industries shows that effective management of AR and overall financial performance of firms is positively correlated. In this paper we address the problem of reducing outstanding receivables through improvements in the collections strategy.”(Zeng et al., 2008)

“We propose an automatic approach to classify invoices into three types: handwritten, machine printed and receipts. The proposed method is based on extracting features using the deep convolution neural network Alex Net” (Tarawneh et al., 2019).

The challenge in this realm involves dealing with complex data and the lack of data related to decisions-making processes not registered in the account receivable system(Appel et al., 2019).

“our aim is to understand customer behavior regarding invoice payments, and propose an analytical approach to learning and predicting payment behavior” (Bahrami et al., 2020)

“This project describes a bag-of-words approach for business invoice recognition. Bags of potential features are generated to capture layout and textual properties for each field of interest, and weighted to reveal key factors that identify a field. Feature selection, threshold tuning, and model comparison are evaluated.”(Wenshun Liu, Billy Wan, n.d.)

Chapter 3: Problem Statement

Customers raise disputes on invoices which eventually delay the invoice payment cycle based on either the disputed invoices get rejected or approved. A typical dispute takes around 6 working days, thus giving the customer an additional 7 days for payment. This is over and above the usual payment term of 30 days. Some disputes are genuine in the sense that invoices might not have met customer requirements as per contractual agreement while booking the orders. However, some customers may take this route to raise false disputes which gave them extra time to pay.

Predicting the disputed status of invoices - will get approved or rejected using appropriate Machine Learning Algorithms and to explore the key drivers leading to disputes. The purpose is to help the business stakeholders to reduce such disputes in future leading to financial stress. Better management of disputes also will lead to better customer satisfaction.

Chapter 4: Objectives of the Study

The scope of this study is to identify/predict the disputed invoices will get rejected or approved. Based on which, rejections can be reduced by taking the positive intervention actions with customers on the disputed invoice and help them to make the payments well within the time and to build the better customer satisfaction in undesirable competitive industry. Can we also identify some latent patterns regarding disputes? Data is mostly in a structured format capturing the lifecycle of disputes. This would form the primary dataset of our study; however, we are not ruling out other related data is required.

Two major objectives of this study is to,

1. *Identify the key features which contributes to the dispute rejections.*
2. *Proactively predict the disputed status of invoice to eliminate delay in payments and improve customer satisfaction.*

Chapter 5: Project Methodology

CRISP-DM framework has been used for this project.

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

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

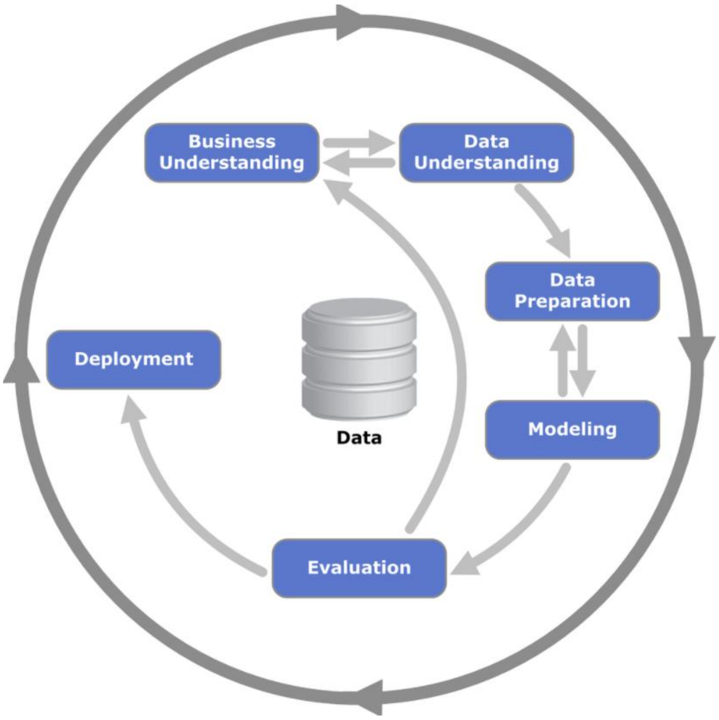


Figure 5.1 “CRISP-DM Framework” (Wikipedia, 2020)

Chapter 6: Business Understanding

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

This project aims to provide algorithmic solutions to the team in predicting the disputed invoice will get either approved or rejected based on the features like reasons for which the invoice is created, who is the requestor, who is approver and in which country is the dispute raised. Based on the proactive predictions number of the rejections can be reduced.

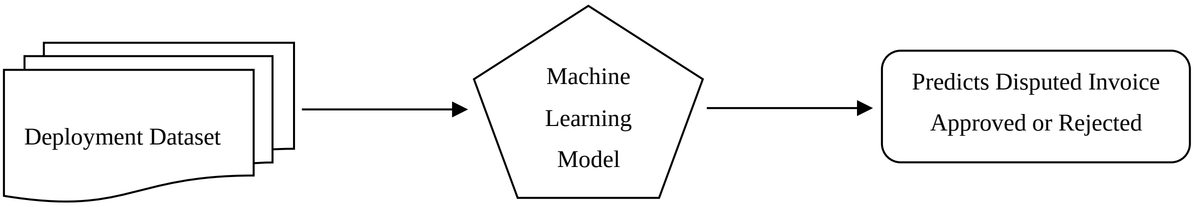


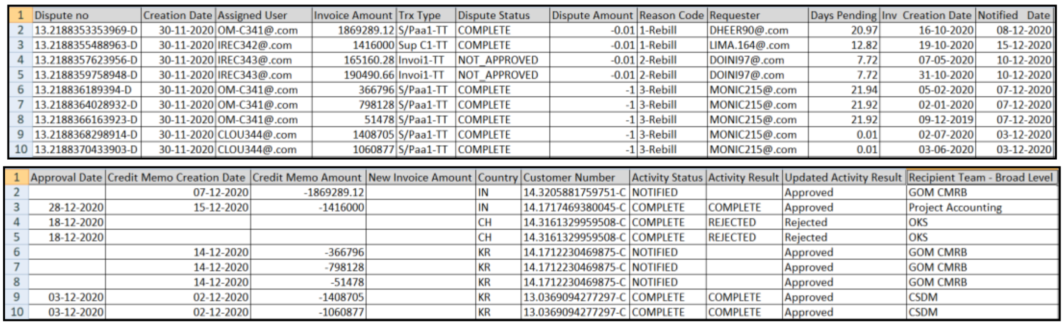
Figure 6.1 “Algorithmic Solution for Disputed Invoice Prediction”

Chapter 7: Data Understanding

Data collected is in the structured format from Company Reports and are in masked format Data is masked so that the organizations and customers confidentially are maintained. Below is the list of features in the given dataset.

Features	Description
Dispute no	System generated number to identify the disputed invoice.
Creation Date	Date on which invoice dispute is created.
Assigned User	Contains email id of the person, who is verifying the dispute.
Invoice Amount	Contain the invoice amount details.
Trx Type	Transaction type of the invoice.
Dispute Status	Contain the different status of the dispute.
Dispute Amount	Contain the invoice amount which is for dispute.
Reason Code	Code for reason for which dispute is raised.
Requester	Contains the email id of the customer.
Days Pending	Days remaining for invoice payment.
Inv Creation Date	Date on which invoice is created.
Notified Date	Date on which disputed is notified.
Approval Date	Date on which disputed invoice gets rejected or approved.
Credit Memo Creation Date	Date on which credit memo is created.
Credit Memo Amount	Memo amount of the credit.
New Invoice Amount	Changed invoice amount.
Country	Country name where dispute is generated.
Customer Number	Unique Identification number of the customer.
Activity Status	High level status of the disputed invoice.
Activity Result	High level status of the disputed invoice which might result into.
Updated Activity Result	Final status of the disputed invoice.
Recipient Team – Board Level	Team for which the dispute is raised.

Table 7.1 Data Dictionary.



1	Dispute no	Creation Date	Assigned User	Invoice Amount	Trx Type	Dispute Status	Dispute Amount	Reason Code	Requester	Days Pending	Inv Creation Date	Notified Date
2	13.2188353353969-D	30-11-2020	OM-C341@.com	1869289.12	S/Paa1-TT	COMPLETE	-0.01	1-Rebill	DHEER90@.com	20.97	16-10-2020	08-12-2020
3	13.2188355488963-D	30-11-2020	IREC342@.com	1416000	Sup C1-TT	COMPLETE	-0.01	1-Rebill	LIMA.164@.com	12.82	19-10-2020	15-12-2020
4	13.2188357623956-D	30-11-2020	IREC343@.com	165160.28	Invoi1-TT	NOT_APPROVED	-0.01	2-Rebill	DOINI97@.com	7.72	07-05-2020	10-12-2020
5	13.2188359758948-D	30-11-2020	IREC343@.com	190490.66	Invoi1-TT	NOT_APPROVED	-0.01	2-Rebill	DOINI97@.com	7.72	31-10-2020	10-12-2020
6	13.218836189394-D	30-11-2020	OM-C341@.com	366796	S/Paa1-TT	COMPLETE	-1	3-Rebill	MONIC215@.com	21.94	05-02-2020	07-12-2020
7	13.2188364028932-D	30-11-2020	OM-C341@.com	798128	S/Paa1-TT	COMPLETE	-1	3-Rebill	MONIC215@.com	21.92	02-01-2020	07-12-2020
8	13.2188366163923-D	30-11-2020	OM-C341@.com	51478	S/Paa1-TT	COMPLETE	-1	3-Rebill	MONIC215@.com	21.92	09-12-2019	07-12-2020
9	13.2188368298914-D	30-11-2020	CLOU344@.com	1408705	S/Paa1-TT	COMPLETE	-1	3-Rebill	MONIC215@.com	0.01	02-07-2020	03-12-2020
10	13.2188370433903-D	30-11-2020	CLOU344@.com	1060877	S/Paa1-TT	COMPLETE	-1	3-Rebill	MONIC215@.com	0.01	03-06-2020	03-12-2020

1	Approval Date	Credit Memo Creation Date	Credit Memo Amount	New Invoice Amount	Country	Customer Number	Activity Status	Activity Result	Updated Activity Result	Recipient Team - Broad Level
2		07-12-2020	-1869289.12		IN	14.3205881759751-C	NOTIFIED		Approved	GOM CMRB
3	28-12-2020	15-12-2020	-1416000		IN	14.1717469380045-C	COMPLETE	COMPLETE	Approved	Project Accounting
4	18-12-2020				CH	14.3161329959508-C	COMPLETE	REJECTED	Rejected	OKS
5	18-12-2020				CH	14.3161329959508-C	COMPLETE	REJECTED	Rejected	OKS
6		14-12-2020	-366796		KR	14.1712230469875-C	NOTIFIED		Approved	GOM CMRB
7		14-12-2020	-798128		KR	14.1712230469875-C	NOTIFIED		Approved	GOM CMRB
8		14-12-2020	-51478		KR	14.1712230469875-C	NOTIFIED		Approved	GOM CMRB
9	03-12-2020	02-12-2020	-1408705		KR	13.0369094277297-C	COMPLETE	COMPLETE	Approved	CSDM
10	03-12-2020	02-12-2020	-1060877		KR	13.0369094277297-C	COMPLETE	COMPLETE	Approved	CSDM

Figure 7.1 Sample Data in masked format.

Exploratory Data Analysis (EDA)

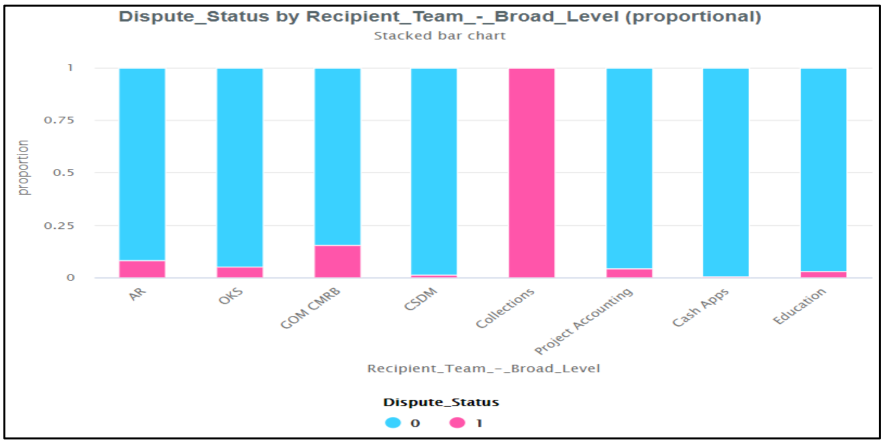


Figure 7.2Stacked Bar Chart – Dispute Status by Recipient Team – Broad Level

Department wise rejections - The disputes are raised from eight departments; Accounts Receivables, Collections, Project Accounting, Cash Applications, Education and others. Among all, the collection team has the highest number of rejects. There is a need to collect the more data on this to understand the existing process and work on to reducing the rejections of the disputed invoices. Cash Apps team has minimum number of rejections on disputed invoices, which also need to be investigated to understand the current process which can be implemented across the other teams to reduce the rejection on disputed invoices.

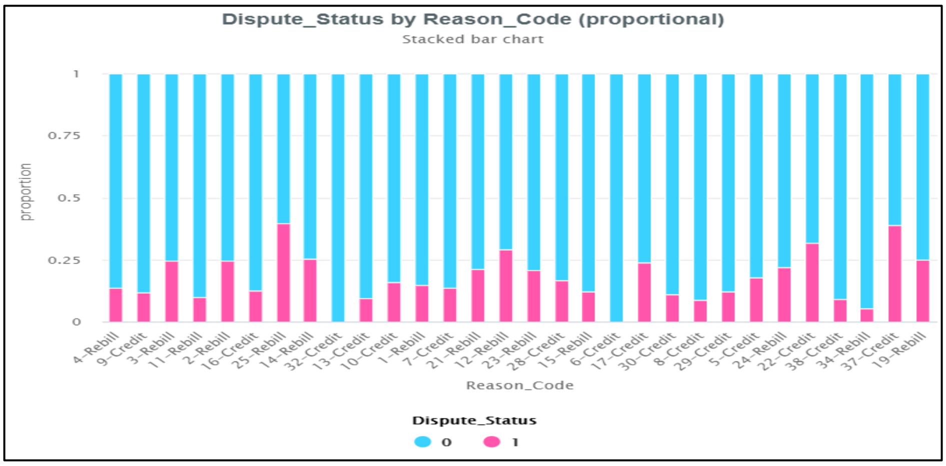


Figure 7.3 Stacked Bar Chart – Dispute Status by Reason Code

Reasons for Disputes: Rebill (Code-25), Credit (Code-37), Credit (Code-22), Rebill (Code-12), Rebill (Code-19) are the reasons codes with more number of rejections. Credit (Code-32), Credit (Code-6) are with minimum number of rejections on disputed invoices. Based on understanding of these reason codes an immediate process improvement can be recommended to bring the rejections of disputes down.

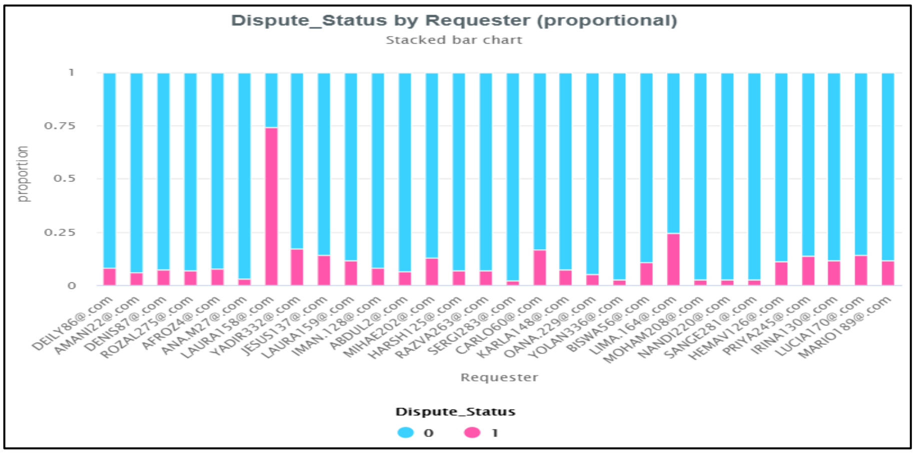


Figure 7.4 Stacked Bar Chart – Dispute Status by Requester

Who Raises Disputes More: LAURA158@.COM, LIMA@.COM and YADIR@.COM are requestor who has the highest number of rejection of disputed invoices. Further understanding is required to connect with these requestors and help them in understanding the SLA’s with organization and if required support them on their business to reduce the disputed invoices.

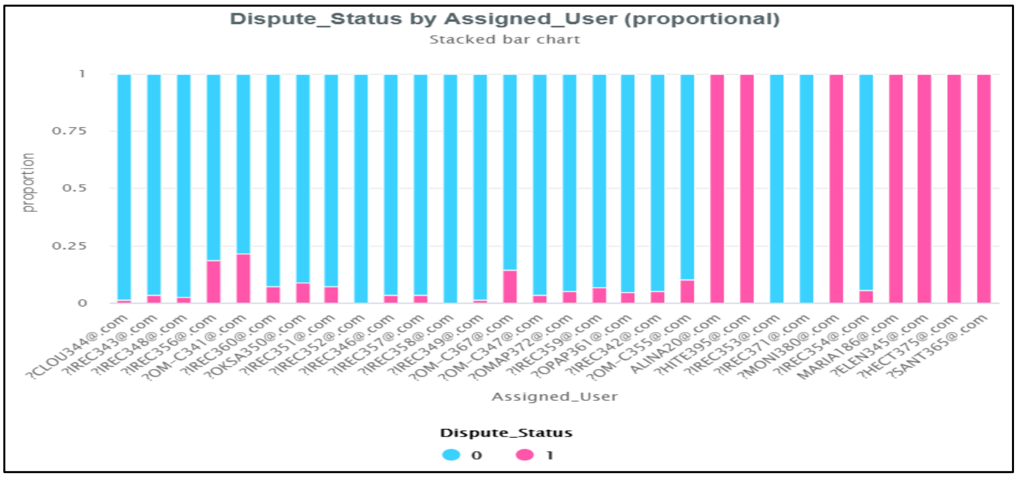


Figure 7.5 Stacked Bar Chart – Dispute Status by Assigned User

Who Rejects the Disputed Invoices: ALINA20@.COM, HITE@.COM, MONI@.COM, MARIA186@.COM, ELEN345@.COM, HECT375@.COM, SANT365@.COM are the assigned users who have rejected the maximum number of disputed invoices. These team members need to be connected further understand why they rejected the disputes. Very high chance for process improvement and support customers in helping them not to raise disputes which have very high chance of rejections.

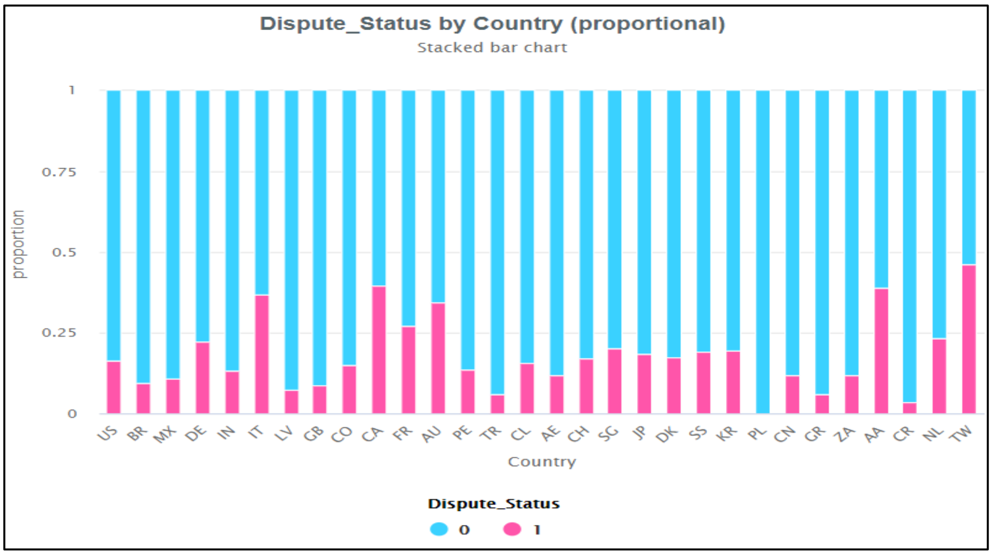


Figure 7.6 Stacked Bar Chart – Dispute Status by Country

Country wise Disputes: Taiwan, Canada, Aruba, Italy and Australia have the highest number of disputed rejections. Further data is required to understand the process which will help organization to improve the process to reduce the rejection of the disputed invoices. Poland and Costa Rica have the minimum number of rejection of disputed invoices understand their working process, so that process can be adapted to other nations on reducing the rejections of disputed invoices.

Chapter 8: Data Preparation

Data is collected from the Database Reports i.e. downloaded in the .csv format. Later data is masked and shared for analysis and study.

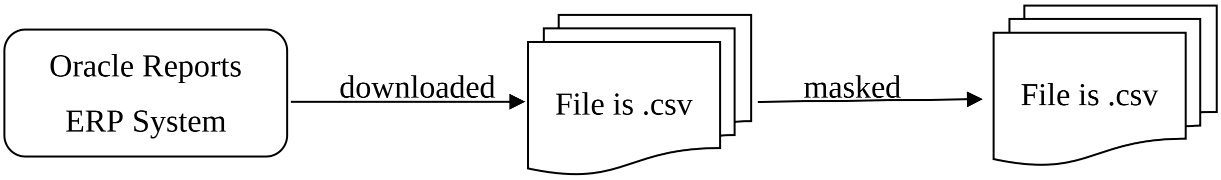
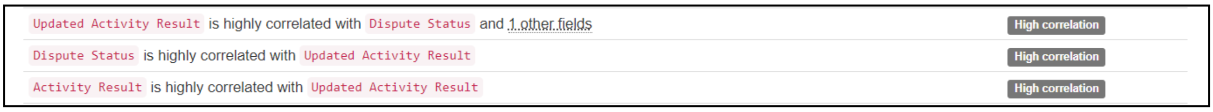


Figure 8.1 Data Extraction Flow



Updated Activity Result is highly correlated with Dispute Status and 1 other fields High correlation

Dispute Status is highly correlated with Updated Activity Result High correlation

Activity Result is highly correlated with Updated Activity Result High correlation

Figure 8.2 EDA Analysis using Pandas Profiling

Test Between	P-Value	Decision : Relationship Between Variables
Dispute_Status and Assigned_User	0.0	Yes
Dispute_Status and Trx_Type	0.00019	Yes
Dispute_Status and Reason_Code	0.00000	Yes
Dispute_Status and Requester	0.00000	Yes
Dispute_Status and Country	0.00000	Yes
Dispute_Status and RecipientTeam_BroadLevel	0.00000	Yes

Table 8.1 Chi-Square Test Analysis

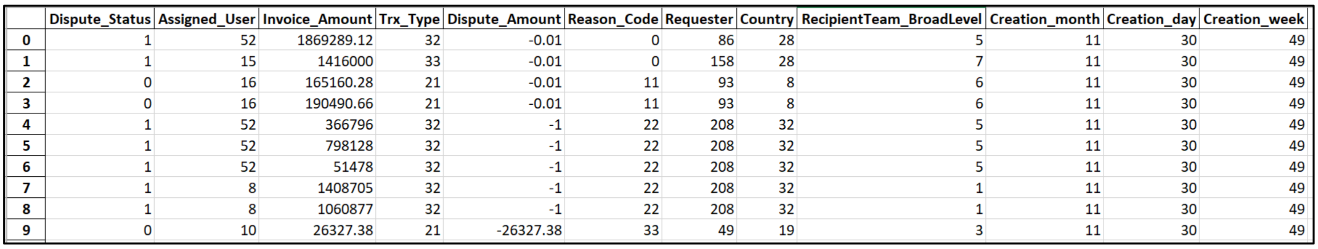
Based on Exploratory Data Analysis (EDA) and the domain experts' suggestions following features have been removed or modified for the further analysis.

Features	**Comments for dropping the features**
Dispute no	Unique Numbers
Customer Number	Unique Numbers for Customers
Updated Activity Result	High correlation and Post factor feature
New Invoice Amount	Post factor feature
Activity Result	High Correlation and Post factor feature
Activity Status	High Correlation and Post factor feature
Approval Date	Post factor feature
Credit Memo Creation Date	Post factor feature
Credit Memo Amount	Post factor feature
Creation Date	Dropped and split the columns into Day, Month and Week
Days Pending	Calculated feature

Table 8.2 Reasons for dropping features

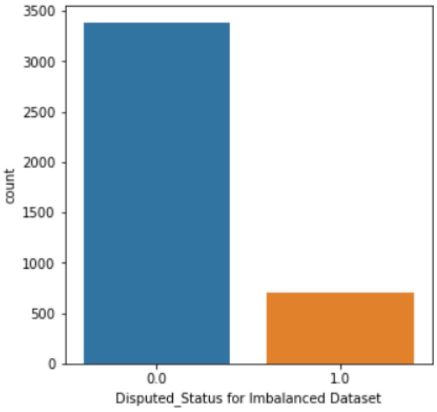
Data Preparation Steps:

- Disputed Status is a categorical feature with the values Complete, Cancelled, Not Approved, Pending Approval and Approved Pending Comp. Based on discussion with domain expert and on the final status of disputed invoice, disputed status is reduced to Approved, Rejected and Pending Approval.
- Dataset is split into three categories training, validation and test set.
- Training and Validation dataset contain disputed status either approved or rejected.
- Test set is used as the deployment data for re-validation of the model.
- Training dataset is imbalanced dataset; it's balanced before building the model.
    - SMOTE packages was used for balancing the data.



	Dispute_Status	Assigned_User	Invoice_Amount	Trx_Type	Dispute_Amount	Reason_Code	Requester	Country	RecipientTeam_BroadLevel	Creation_month	Creation_day	Creation_week
0	1	52	1869289.12	32	-0.01	0	86	28	5	11	30	49
1	1	15	1416000	33	-0.01	0	158	28	7	11	30	49
2	0	16	165160.28	21	-0.01	11	93	8	6	11	30	49
3	0	16	190490.66	21	-0.01	11	93	8	6	11	30	49
4	1	52	366796	32	-1	22	208	32	5	11	30	49
5	1	52	798128	32	-1	22	208	32	5	11	30	49
6	1	52	51478	32	-1	22	208	32	5	11	30	49
7	1	8	1408705	32	-1	22	208	32	1	11	30	49
8	1	8	1060877	32	-1	22	208	32	1	11	30	49
9	0	10	26327.38	21	-26327.38	33	49	19	3	11	30	49

Figure 8.3 Sample data after preprocessed



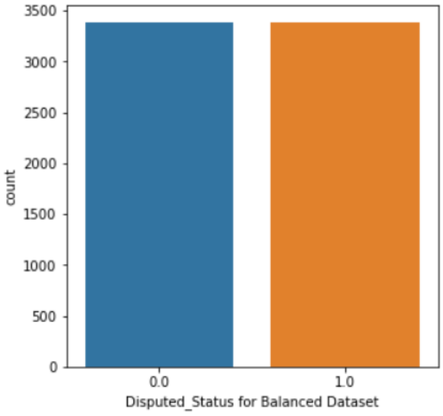


Figure 8.4 Bar Graph for Imbalanced Dataset

Figure 8.5 Bar Graph for Balanced Dataset

Imbalanced Dataset: Disputed status has the very high difference between the approved and rejected status values. Thus we say that the dataset is imbalanced. Figure 8.4.

Balanced Dataset: Disputed status have approximately or same number of approved or rejected status values. Then we say dataset is balanced. Figure 8.5.

Once the dataset was balanced the count of the rejected values were almost equal to the approved values. This was achieved by SMOTE package in python.

Chapter 9: Data Modeling

Preprocessed data discussed in the previous section was fed into to multiple models to get the predicted values of disputed invoices.

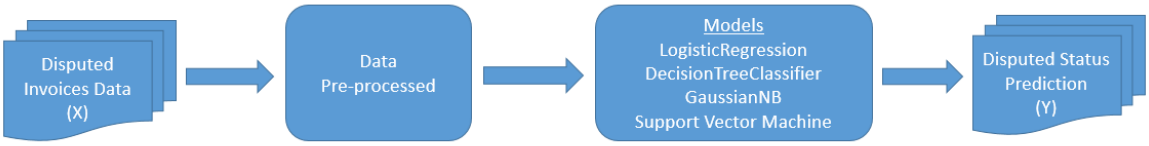
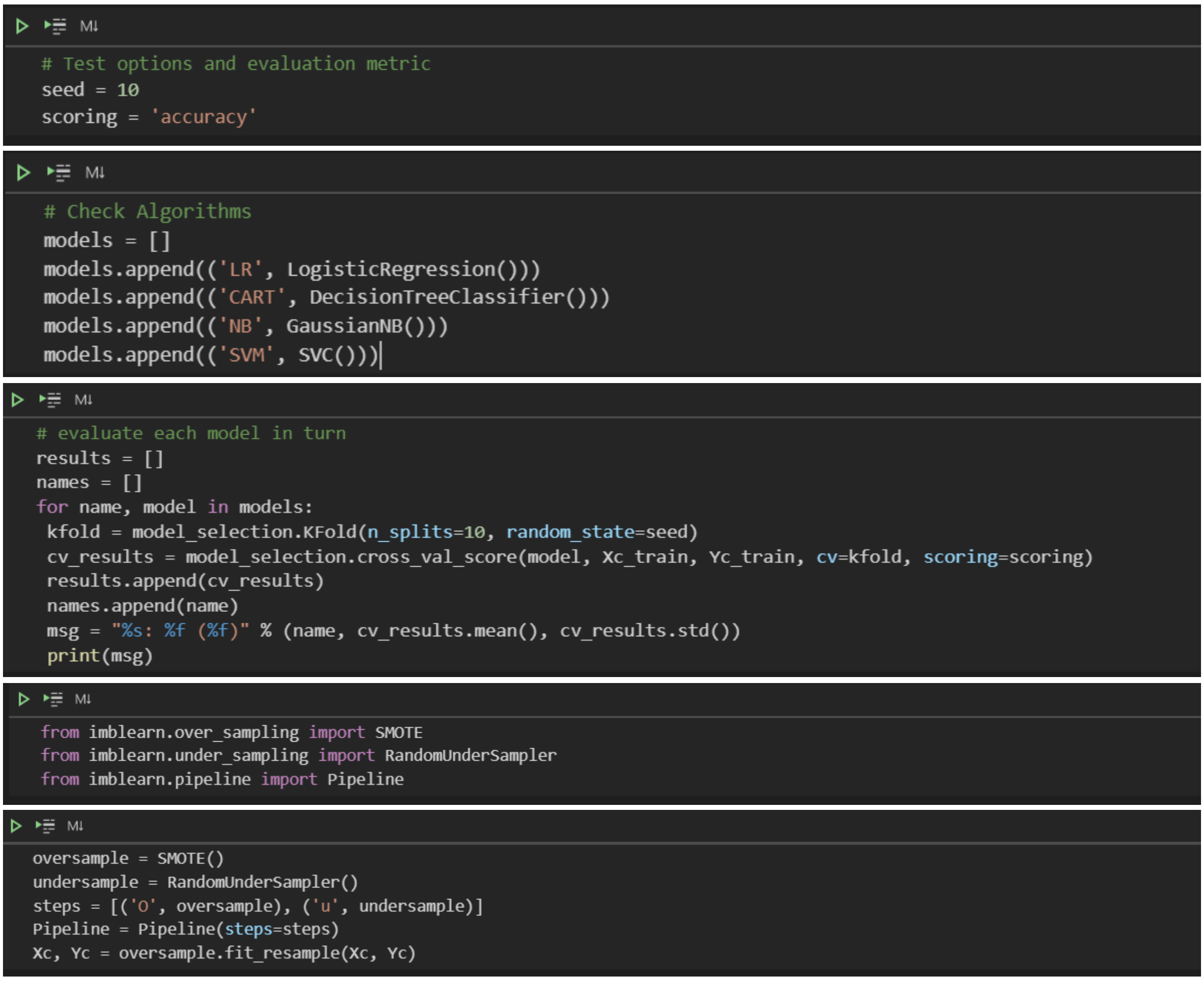


Figure 9.1 Pre-Proceed data into Model Flow.

Classification algorithm, a supervised learning technique used to identify / predict the categorical observations on basis of training set. Program learns from the given data and classifies it into classes or groups. Four classification techniques have been used; Logistic Regression, Decisions Tree Classifier, Support Vector Classifier and Naïve Bayes Model as shown below.



```
# Test options and evaluation metric
seed = 10
scoring = 'accuracy'
```

```
# Check Algorithms
models = []
models.append(('LR', LogisticRegression()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))
```

```
# evaluate each model in turn
results = []
names = []
for name, model in models:
 kfold = model_selection.KFold(n_splits=10, random_state=seed)
 cv_results = model_selection.cross_val_score(model, Xc_train, Yc_train, cv=kfold, scoring=scoring)
 results.append(cv_results)
 names.append(name)
 msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
 print(msg)
```

```
from imblearn.over_sampling import SMOTE
from imblearn.under_sampling import RandomUnderSampler
from imblearn.pipeline import Pipeline
```

```
oversample = SMOTE()
undersample = RandomUnderSampler()
steps = [('O', oversample), ('u', undersample)]
Pipeline = Pipeline(steps=steps)
Xc, Yc = oversample.fit_resample(Xc, Yc)
```

Figure 9.2 Models used for Predictions

Chapter 10: Data Evaluation

Accuracy of the models of Logistic Regression (LR), Decision Tree Classifier (CART), Naïve Bayes (NB) and Support Vector Machine (SVM) are as follows:

Models	Accuracy
Logistic Regression	89.63%
Decision Tree Classifier	93.45%
GaussianNB	88.53%
Support Vector Machine	90.09%

Table 10.1 Metrics for Imbalanced data

Models	Accuracy
Logistic Regression	46.26%
Decision Tree Classifier	96.11%
GaussianNB	49.58%
Support Vector Machine	49.79%

Table 10.2 Metrics for Balanced Data

Decision Tree Classifier (CART) Model has an accuracy value of around 94% which best suits the data for predicting the disputed invoices will get approved or rejected when compared to other models used for modeling.

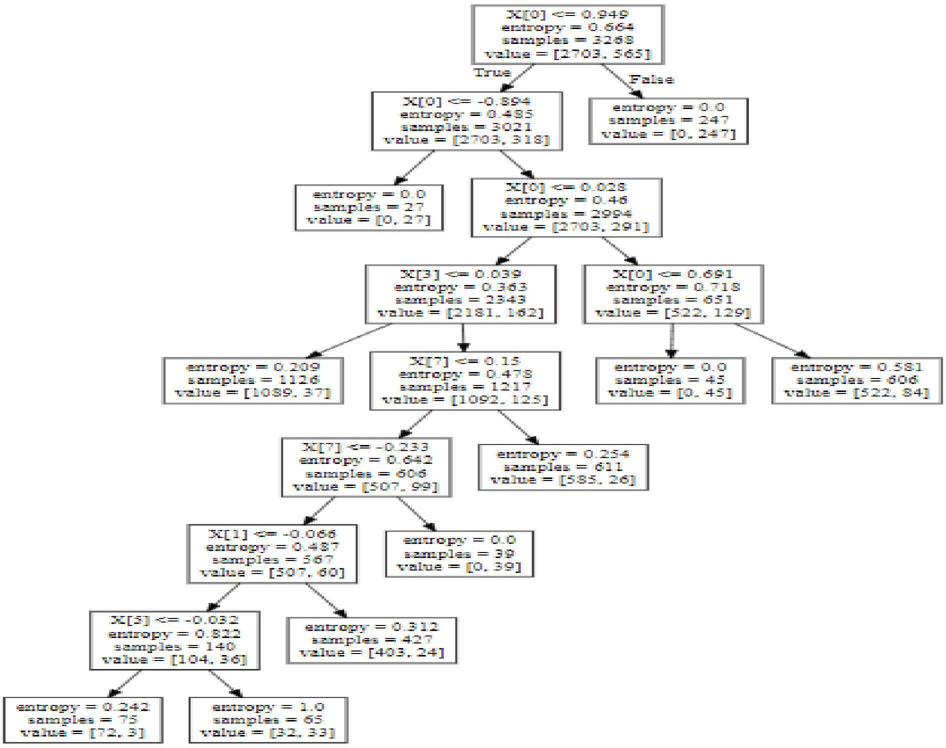


Figure 10.1 Decision Tree with parameter of max_leaf_node of 10

From the above Decision Tree, we see that X [0] points to the feature Assigned User which is used for 1st split with an entropy of 0.664 for that branch, then followed by X [3] Dispute Amount, X [7] Recipient Team – Board Team, X [1] Invoice Amount and then followed by X [5] Requestor.

Failure of classification Accuracy for Imbalanced Class Distributions (MachineLearningMastery, n.d.)

Chapter 11: Deployment

Test dataset which had the third category of dispute status of value 2 or approval pending was used for the prediction of dispute status will get approved or rejected. Prediction was done on the balanced dataset.

Row Labels	**Count of Predicted_Dispute_status**
0	338
1	577
Grand Total	**915**

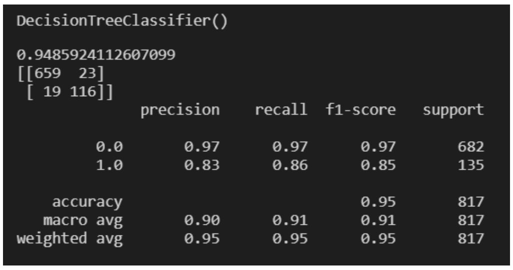
Table 11.1 Prediction values of approved (0) and rejected (1) values of disputed invoice.

Out of 915 disputed invoices338 were rejected and 577 were approved.

Dispute status value will be validated against the actual status of disputed invoice number from the organizations based on which next actions will be taken for the implementation of the model.

Chapter 12: Analysis and Results

“Classification Accuracy: it defines how often the model predicts the correct output. It can be calculated as the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifier” (Javapoint-ConfusionMatrix, 2020)



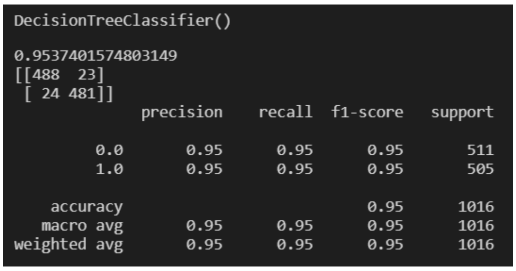
```
DecisionTreeClassifier()

0.9485924112607099
[[659  23]
 [ 19 116]]
              precision    recall  f1-score   support

         0.0       0.97      0.97      0.97       682
         1.0       0.83      0.86      0.85       135

    accuracy                           0.95       817
   macro avg       0.90      0.91      0.91       817
weighted avg       0.95      0.95      0.95       817
```

Figure 12.1 Metrics for imbalanced data.



```
DecisionTreeClassifier()

0.9537401574803149
[[488  23]
 [ 24 481]]
              precision    recall  f1-score   support

         0.0       0.95      0.95      0.95       511
         1.0       0.95      0.95      0.95       505

    accuracy                           0.95      1016
   macro avg       0.95      0.95      0.95      1016
weighted avg       0.95      0.95      0.95      1016
```

Figure 12.2 Metrics for balanced data.

Confusion Matrix: Model was able to identify 659 approved statuses correctly, with 23 approved statuses was identified wrongly. 116 rejected statuses were identified correctly with 19 rejected statuses identified wrongly from the imbalanced data.

Precision value for imbalanced and balanced is around 0.95 for predicting the approved status and 0.83 for unbalanced, 0.95 to the balanced data, i.e. the performance of the model can correctly predict the true positive values from the false positive values for the approved dispute status. Value closer to 1 is the better model.

Recall metric value for imbalanced and balanced is around 0.95 for approved status and 0.86 for unbalanced and 0.95 for balanced data, i.e. the performance of the model can predict the true positive from total positive values for the approved dispute status. Value closer to 1 is the better model.

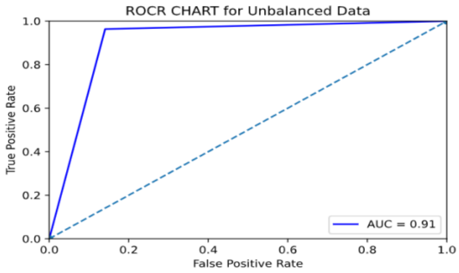


Figure 12.3 ROC graph for imbalanced Data

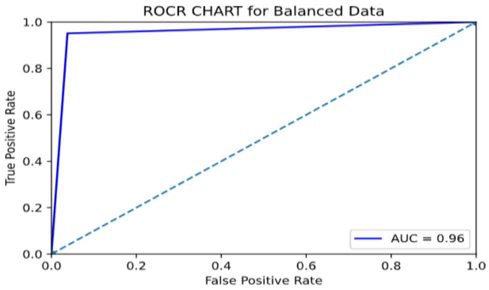


Figure 12.4 ROC graph for Balanced Data

AUC in ROC graph is closer to top left corner indicting the better performance. This metric holds good for the balanced dataset, not preferred for the imbalanced data.

Chapter 13: Conclusions and Recommendations for future work

This project solely focuses on the predicting the disputed invoices will get approved or rejected based on ML model, where the model studies the historical data and reduce the number of disputed invoices rejections and make the customers pay the invoice well with in time.

Recommendations for further work: This project does not cover why these disputed invoices gets created. At a high level the data shows the following:

- Who are the customers who regularly create these disputes?
- Which countries maximum number of invoices which are disputed,
- The reasons quoted for disputes etc.

Further collection of data and detailed analysis is required at different team, on different reason codes, assigned users and process and SLA's of different countries, there is a high chance on improving the process and reducing the rejection of disputed invoices. This will help in reducing the cost on intervention actions between the internal teams and customers and to build an achieve better customer satisfaction.

Bibliography

Appel, A. P., Oliveira, V., Lima, B., Malfatti, G. L., de Santana, V. F., & de Paula, R. (2019). Optimize cash collection: Use machine learning to predicting invoice payment. *ArXiv*.

Bahrami, M., Bozkaya, B., & Balcisoy, S. (2020). Using Behavioral Analytics to Predict Customer Invoice Payment. *Big Data*, *8*(1), 25–37. https://doi.org/10.1089/big.2018.0116

Cheong, M. L. F., Cheong, M. L. F., & Shi, W. (2018). Customer level predictive modeling for accounts receivable to reduce intervention actions Customer Level Predictive Modeling for Accounts Receivable to Reduce Intervention Actions. *Proceedings of the 14th International Conference on Data Science (ICDATA 2018), Las Vegas, Nevada, July 30 - August 2.*, *Icdata*, Research Collection School Of Information Systems.

Fernandez, E. B., & Yuan, X. (2010). An analysis pattern for invoice processing. *ACM International Conference Proceeding Series*, *August 2009*. https://doi.org/10.1145/1943226.1943239

Hu, W. (2009). *Overdue Invoice Forecasting and Data Mining*.

Javapoint-ConfusionMatrix. (2020). *Confusion Matrix in Machine Learning - Javatpoint*. https://www.javatpoint.com/confusion-matrix-in-machine-learning

MachineLearningMastery. (n.d.). *Failure of Classification Accuracy for Imbalanced Class Distributions*. Retrieved March 3, 2021, from https://machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/

Peiguang, H. (2015). Predicting and Improving Invoice-to-Cash Collection Through Machine Learning. *Master Thesis*, 1–92. http://dspace.mit.edu/bitstream/handle/1721.1/99584/925473704-MIT.pdf?sequence=1

Shah, H. S. (2019). Customer Payment Prediction in Account Receivable. *International Journal of Science and Research (IJSR)*, *8*(1), 642–644. https://www.ijsr.net/archive/v8i1/ART20194177.pdf

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

Tater, T., Dechu, S., Mani, S., & Maurya, C. (2018). Prediction of invoice payment status in account payable business process. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11236 LNCS*(September 2019), 165–180. https://doi.org/10.1007/978-3-030-03596-9_11

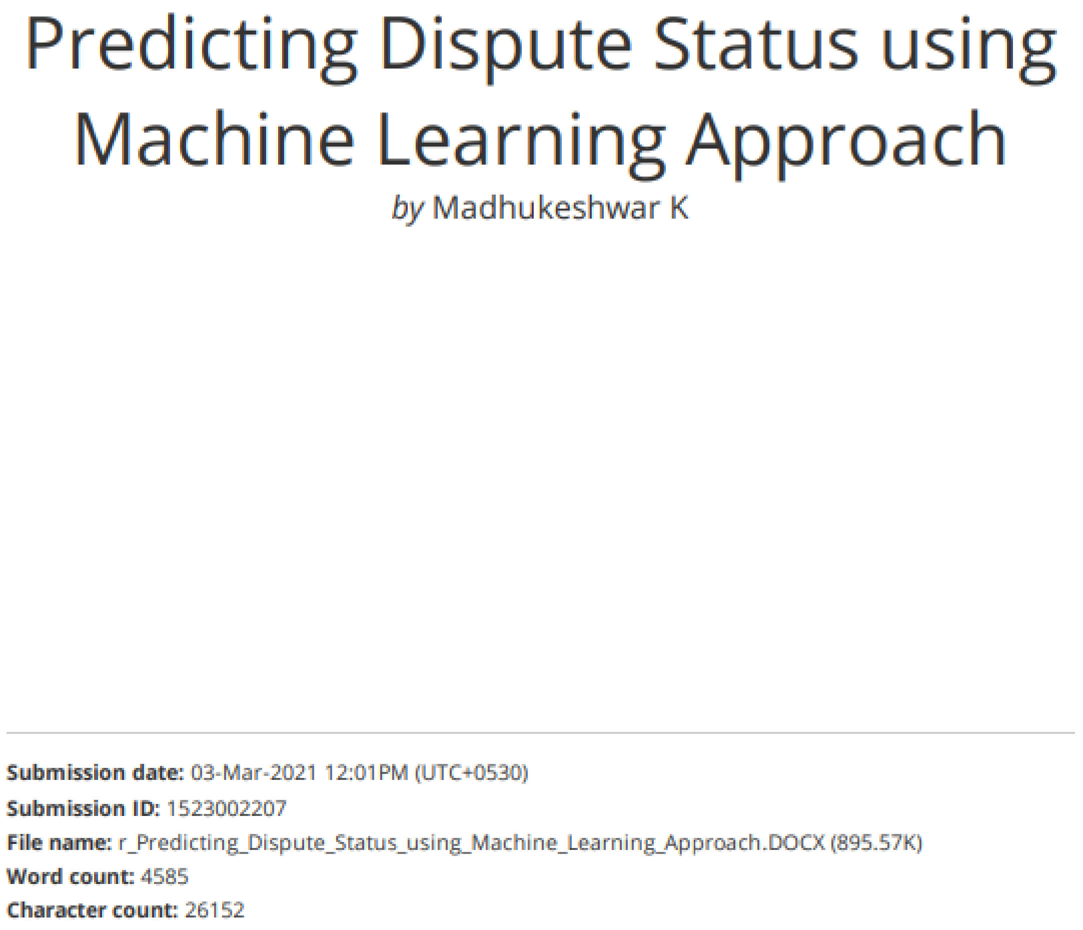
Wenshun Liu, Billy Wan, Y. Z. (n.d.). Unstructured Document Recognition on Business Invoice. *Stanford ITunes University*.

Wikipedia. (2020). *Cross-industry standard process for data mining - Wikipedia*. https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Zeng, S., Melville, P., Lang, C. A., Boier-Martin, I., & Murphy, C. (2008). Using predictive analysis to improve invoice-to-cash collection. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *October 2014*, 1043–1050. https://doi.org/10.1145/1401890.1402014

Appendix

Plagiarism Report[1]



Predicting Dispute Status using Machine Learning Approach

by Madhukeshwar K

Submission date: 03-Mar-2021 12:01PM (UTC+0530)

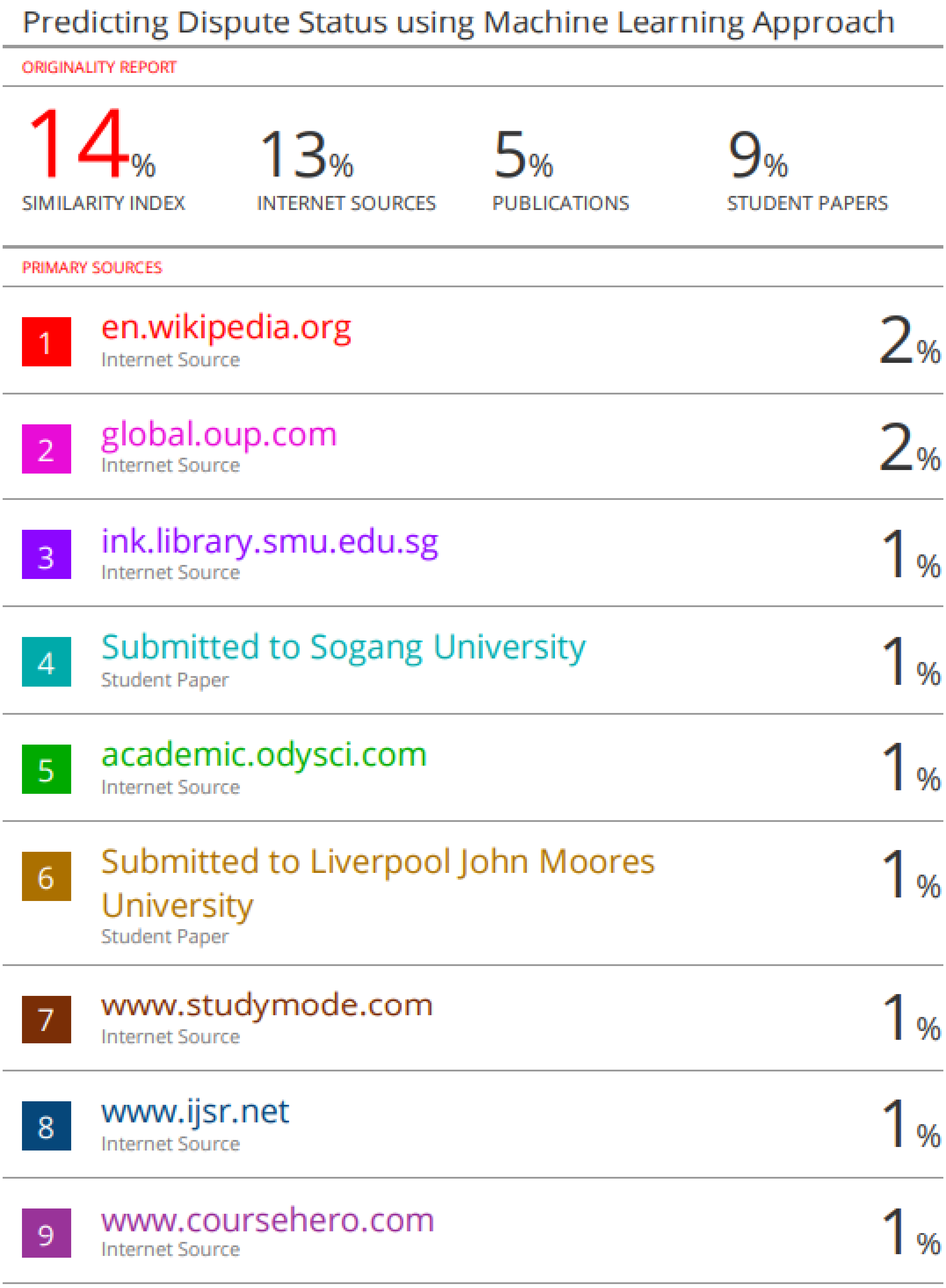
Submission ID: 1523002207

File name: r_Predicting_Dispute_Status_using_Machine_Learning_Approach.DOCX (895.57K)

Word count: 4585

Character count: 26152

[1] Turntn report to be attached from the University.



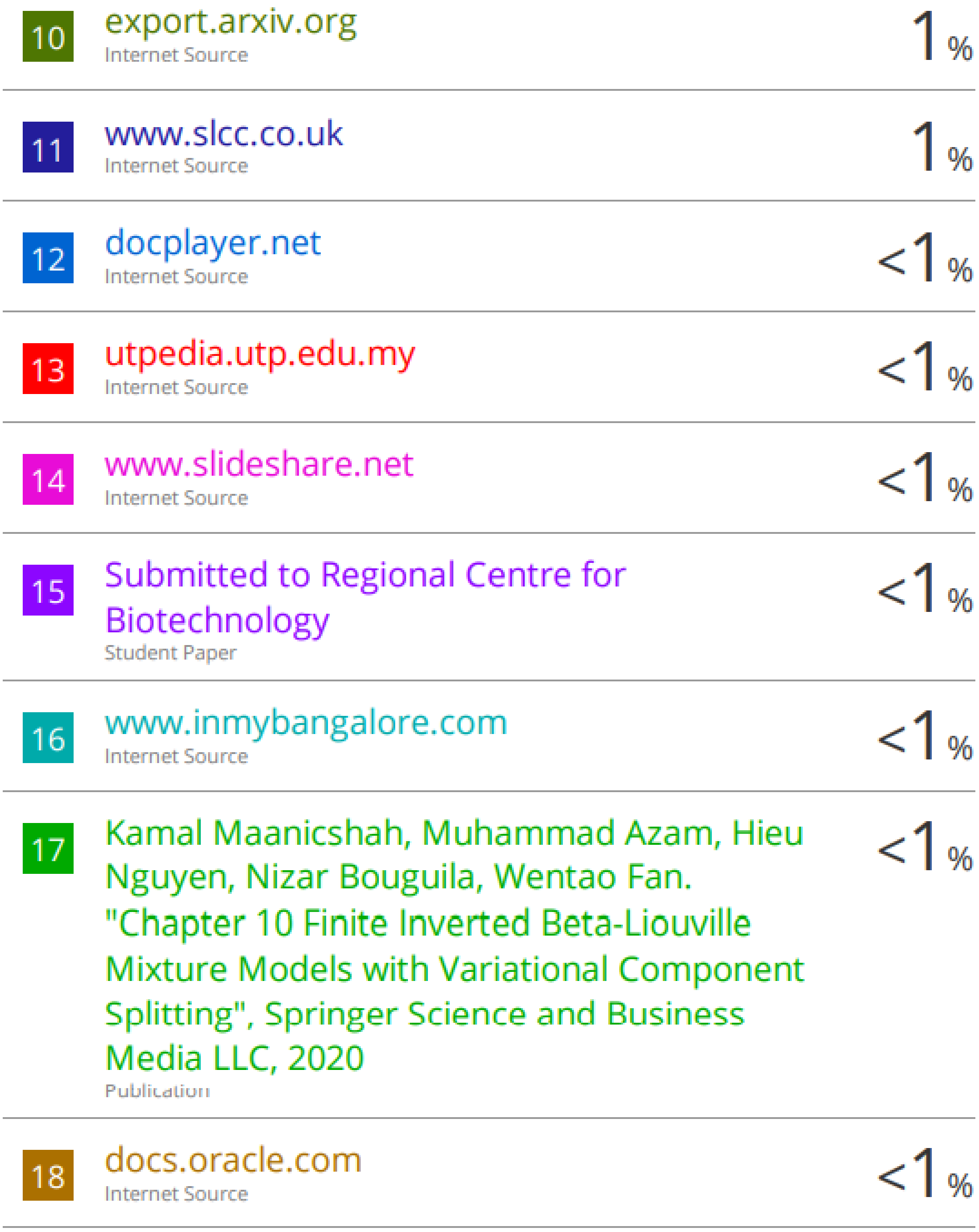
Predicting Dispute Status using Machine Learning Approach

ORIGINALITY REPORT

14%	13%	5%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	en.wikipedia.org Internet Source	2%
2	global.oup.com Internet Source	2%
3	ink.library.smu.edu.sg Internet Source	1%
4	Submitted to Sogang University Student Paper	1%
5	academic.odysci.com Internet Source	1%
6	Submitted to Liverpool John Moores University Student Paper	1%
7	www.studymode.com Internet Source	1%
8	www.ijsr.net Internet Source	1%
9	www.coursehero.com Internet Source	1%



10	export.arxiv.org Internet Source	1 %
11	www.slcc.co.uk Internet Source	1 %
12	docplayer.net Internet Source	<1 %
13	utpedia.utp.edu.my Internet Source	<1 %
14	www.slideshare.net Internet Source	<1 %
15	Submitted to Regional Centre for Biotechnology Student Paper	<1 %
16	www.inmybangalore.com Internet Source	<1 %
17	Kamal Maanicshah, Muhammad Azam, Hieu Nguyen, Nizar Bouguila, Wentao Fan. "Chapter 10 Finite Inverted Beta-Liouville Mixture Models with Variational Component Splitting", Springer Science and Business Media LLC, 2020 Publication	<1 %
18	docs.oracle.com Internet Source	<1 %

19	Submitted to Visvesvaraya Technological University, Belagavi Student Paper	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

GitHub

https://github.com/kyasanur/Predicting-Dispute-Status-using-Machine-Learning-Approach