Invoice Deduction Classification Using LGBM Prediction Model

Laharika Tutica, KSK Vineel, Sushruta Mishra, Manoj Kumar Mishra, and Saurabh Suman

Abstract Deductions are predominantly the short payments done for a generated invoice usually by the customer as a compensation or for the lack of products or services. Possible reasons for deductions to happen include shortage, damage, late delivery, and other-related factors. The machine learning approach has a huge impact on the deduction domain as eliminates the manual effort of a deduction analyst without compromising much on the accuracy. A deduction analyst can save so much on time as now he/she does not have to go through the complex procedure of deduction validity or invalidity. Also this solution will help in speeding up the business process which will lead to customer satisfaction due to on-time delivery. In this research, various machine learning techniques like LGBM and random forest are used for the analysis. It was observed that LGBM model provided optimum result thereby helping business analysts to take decision with respect to invoice payments.

Keywords Deductions · Machine learning · LGBM · Random forest

L. Tutica · K. Vineel · S. Mishra (✉) · M. K. Mishra · S. Suman
School of Computer Engineering, KIIT (Deemed to be University), Bhubaneswar, India
e-mail: mishra.sushruta@gmail.com

L. Tutica
e-mail: laharikatutica@gmail.com

K. Vineel
e-mail: 1628076@kiit.ac.in

M. K. Mishra
e-mail: manojku.mishra05@gmail.com

S. Suman
e-mail: saurabhsu.jmp@gmail.com

P. K. Mallick et al. (eds.), *Advances in Electronics, Communication and Computing*,
Lecture Notes in Electrical Engineering 709,
https://doi.org/10.1007/978-981-15-8752-8_13

1 Introduction

Deductions are predominantly the short payments done for a generated invoice usually by the customer as a compensation or for the lack of products or services. For example, when a customer does not make a full payment to the client as much as the actual amount present in the generated invoice (also called an open invoice), a deduction results. Deductions are also known as dispute or chargebacks; intrinsically, these terms are used in the industry but all are meant to be identical. An example will enlighten us more about the deduction.

There are possibilities of two cases for the cause of deductions which are listed below.

Case 1: An invoice of 100$ is raised by the company *X* (let's assume) to a company *Y* (let's assume) and a payment of 80$ is made by the *Y* to *X* then there comes a difference of 20$ which will result in deduction.

Case 2: An invoice is raised by the company *X* to the company *Y* of 1000$ and due to some discrepancy *Y* do not want to pay any amount (due to some reasons the company does not pay the full amount) then company *Y* will wait for the company *X* to issue credit memo document (alias: credit invoice) worth of 300$ and then the payment is processed for rest 700$.

Fig. 1 depicts an overview of The Deductions Cloud. It is divided into 3 phases, i.e., the company(seller), the work flow and collaboration engine(internet/cloud) and the customer(buyer/payer).

The ERP is the main source where the whole data is stored and accessed by the different organization in a better organized way to do the analysis on variety of data received, the ERP systems basically integrate all the data and process of

Fig. 1 Deduction cloud overview

the organization into one unified system and also helps to automate the business functions.

Deductions cloud enables a proactive deduction management operation, i.e.,

- The solution streamlines processing
- Shortens resolution cycle time
- Reduces processing costs
- Increases recovery rates on invalid deductions
- Provides automation
- Process standardization
- A platform for cross-departmental and customer collaboration.

The processes which are being held in the cloud is to get the clean and cured data which will make a user for the better understanding of the validity of the deduction which is the ultimate result or the solution to be achieved. The cloud also provides the foremost robust automation engine available to capture deduction data from customers and provide the knowledge required for resolution. Backup documentation, like proof of delivery (PODs), bill of lading (BOLs) are captured automatically and linked to the corresponding deductions to scale back manual research. Corresponding trade promotions also are identified and suggested for settlement.

1.1 Reasons Behind Deduction Occurrence

Possible reasons for deductions to happen include the following:

- Shortage: Insufficient amount of goods in the inventory resulting in less number of goods delivered to the customer.
- Damage: Delivery of damaged or defective goods to the customer.
- Wrong products being delivered: Goods might get exchanged or wrongly delivered due to human error.
- Late delivery: Delayed delivery of goods than the expected date.
- Ongoing promotion: Reduction in the price of the goods in order to promote their product or due to early payments by the customer to avail a discount offered by the seller.

One deduction analyzing and dissecting each and every deduction being made to be a valid or an invalid one and also furnishing with apt solutions for every valid deduction is humanly very time-consuming and difficult since the data size is very huge.

Hence, the solution impact on solving the deductions problem using machine learning algorithms is:

- The machine learning approach has a huge impact on the deduction domain as eliminates the manual effort of a deduction analyst without compromising much on the accuracy.

Table 1 Data insights of the invoice data used

Account used	Company (provided by the organization to classify the deductions)	
Dimension	(136,530 × 26)	
Class	0	Invalid deduction
	1	Valid deduction
Class distribution	Invalid deduction (Class 0)	1947 records
	Valid deduction (Class 1)	134,583 records
Invalid deduction ratio	2.7% of the total records	
Features used	['original dispute amount', 'company code', 'deduction created month', 'ar reason code']	
Output label	['fk action code id'] and ['correspondence flag'] **Action code = "Denied to Customer-840, Refused to Pay-843", Correspondence_flag = 1; then that deduction is considered as invalid. (Just an illustration for better understanding.)	
Training data size	Data from January, 1, 2016 to June, 30, 2017 [124,258 records]	
Testing data size	Data from July, 1, 2017 to February, 2, 2018 [12,272 records]	

- A deduction analyst can save so much on time as now he/she does not have to go through the complex procedure of deduction validity/invalidity.
- Many deductions can be sorted out in short span of time effortlessly.
- Also the client can save money by employing less number of deduction analysts for research process, instead those analysts can be put into some other useful work.
- Also this solution will help in speeding up the business process which will lead to Customer Satisfaction due to on-time delivery.

2 Data Insights

See table 1.

3 Explanation of Output Variable

Since it is a problem of supervised binary classification where output variable primarily consists of either '0' or '1' as the integer values by corresponding to the invalid deduction and valid deduction, respectively.

Output Label: Class 0: Invalid deduction
Class 1: Valid deduction

```
df['is_valid'] = np.where((df['fk_action_code_id'].isin([840,843])|(df['correspondence_flag']==1)),0,1)
```

An output variable is created according to the conditions accorded here, the following condition for the deduction to be invalid was Action code = 'Denied to Customer-840, Refused to Pay-843' and 'Correspondence_flag = 1'.

4 Related Work

Smirnov [1] summarized that random survival forests model, which additionally uses historical payment behavior of debtors, performs better in ranking payment times of late invoices than traditional Cox Proportional Hazards model. Tater et al. [2] propose a different approach to the matter and instead of predicting invoice in accounts receivable, they particularize in accounts payable, working on invoices that were already delayed. Similarly, Younes [3] focuses on accounts payable case and attempts to deal with the difficulty of invoice processing time interval, understanding the overdue invoices, and thus the impact of delays within the invoice processing. Invoice payment prediction could even be modeled as a classification problem, but there is just a little body of work that addresses this problem. One of the few works that investigate this is often Zeng [4], where the authors formulate the matter as traditional supervised classification and apply existing classifiers thereto. Dirick et al. [5] tested several survival analysis techniques in credit data from Belgian and Great Britain financial institutions. The matter of predicting invoice payment has been traditionally tackled using statistical survival analysis methods, such as the proportional hazards method [6]. Sushruta et al. [7] developed a resampling-based pre-processing technique to deal with the skewing of unbalanced datasets and classified various sorts of tumor in patients. Soumya et al. [8] used LVQ technique for instance and analyze the clustering deviation issue on carcinoma dataset.

5 Machine Learning Algorithms Used in Study

In this study, some popular and efficient machine learning algorithms are used for binary classification of deductions for invoice which include logistic regression, random forest, and LGBM classifier.

5.1 Logistic Regression

Logistic regression looks almost like rectilinear regression which is borrowed from the sector of statistics like many other machine learning models despite having a

reputation like regression it is not used for predicting the continual values(like height, weight, etc.) rather it is used for the binary classification(like 0 or 1, Yes or No, etc.) problems. This method fits an S-shaped logistic function, i.e., the Sigmoid-Function which is an S-shaped curve which will take any real-valued number and map it into a worth between the range of 0 and 1, but never exactly at those limits. These values between 0 and 1 will then be transformed into either class 0 or 1 employing a threshold classifier. Its ability to supply probabilities and classify new samples using continuous and discrete measurements makes it a 1 of the favored machine learning method. But sometimes, using this system may cause overfitting. And also it gives low predictive performance which is why, despite being a really popular machine learning technique, it is not always relied upon.

5.2 Random Forest

The random forest can work for classification and as well as regression problems. This machine learning algorithm helps to select the required output from the decision trees (Capable to work for both categorical and continuous input and discrete variables) which is the base class of the random forest. This uses a technique of splitting into n number of trees and decides by selecting one class out of it. The multiple trees chooses the classification having the most votes (over all trees of the forest) and when it comes to the regression, the forest takes out the average of outputs of different trees. This learning method has a power of handling large data sets with higher dimensionality. But since this machine learning method generates n number of trees, it needs more computational power and resources which is not cost efficient. The train data set takes longer training time and hence it is not time efficient.

5.3 Light Gradient Boosting Machine (LGBM)

Boosting is an example of ensemble learning and a technique where we convert a set of all the weak learners into strong ones. LGBM is a gradient boosting technique that uses tree-based learning algorithms.

It is designed to be distributed and efficient with the following advantages:

- Faster training speed and higher efficiency.
- Lower usage of memory and comparatively better accuracy.
- Support of parallel and graphical processing unit learning.
- Capable of handling huge-scale data.

This algorithm also helps us to find out the regression problems (for numeric outcomes) and also to find out the classification problems (for categorical outcomes).

Boosting helps to give good weight-age to the data so that we can find some hidden inferences from the data. Gradient boosting is a specialized type of boosting framework—where it performs on reducing error subsequently.

6 Proposed Work

The Workflow performs the following tasks (Fig. 2):

6.1 Data Extraction and Pre-processing of the Invoice Data

The raw data was extracted and acquired from the central database, known as data acquisition. After extraction, the acquired invoice data was pre-processed, i.e., to handle the various missing data, outliers present, and the long tails (data at the extreme ends of the domain range). Once the pre-processing is done, data transformation takes place which involves enriching and standardizing the data, converting the data format compatible with the training algorithms. Later, train-test splitting is performed. Since the entire data cannot be used to train the model which would lead to overfitting of the model and result in wrong predictions for the unknown data with less accuracy rates, the data set is usually split into train data and test data where the test data is kept hidden from the model while training the model.

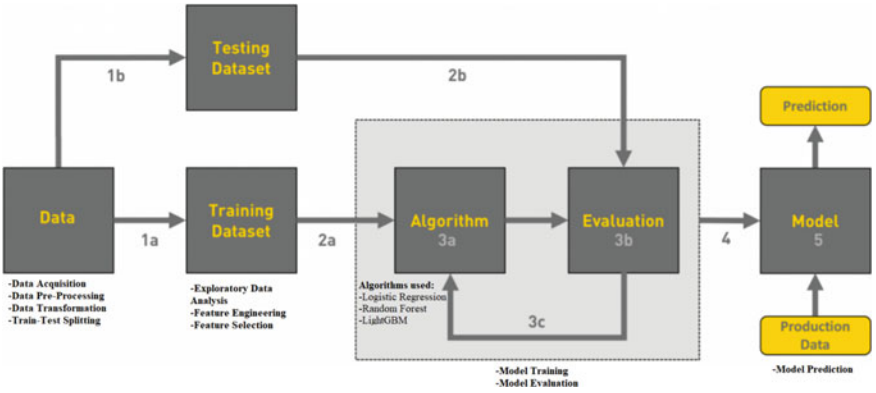

Fig. 2 Workflow of the proposed work

6.2 Training the Invoice Data

The EDA is performed on the train data set to know the past nature. It helps to gain some inferences and insights from the data set and after the EDA, so that, we can have assumptions required for model fit and testing purposes, we can also transform the variables accordingly as and when required. The process of creating new features from the existing columns is called feature engineering. The new features should be created in such a manner that they should have a large impact on the data set to give a good accuracy and metric scores. The features which are taken into the final set of features is called feature selection because the feature importance of those features is good so we should select the features wisely which could help to give a better accuracy rate despite being in a less number.

6.3 Algorithm and Evaluation Using Various Machine Learning Algorithms

Based on the features selected and the problem statement given, various machine learning algorithms could be implied. The algorithm which results in giving the best accuracy and precise metric scores is chosen to be the final technique and the train dataset is trained and evaluated using that particular algorithmic technique. In this proposed work, various machine learning algorithms like logistic regression, random forest, and LightGBM were applied and the best results were found out to be using the LightGBM technique.

6.4 Model Prediction Using the Best Suited Machine Learning Algorithm

Once the machine learning algorithm to be used is finalized based on the recall and precision results calculated, the model is provided with the test data set which is kept hidden from the model while training and the accuracy rate is calculated.

The noted accuracy rate, precision, and recall scores calculated will be the final results obtained.

7 Result Analysis

The system model was implemented with our machine learning classifiers. There are several performance metrics that may be considered for valid and invalid deduction prediction.

Table 2 Precision and recall comparison analysis

Techniques	Class	Precision	Recall
Logistic regression	0	0.03	0.63
	1	0.99	0.74
Random forest classifier	0	0.05	0.82
	1	1.00	0.82
Light Gradient boosting machine	0	0.07	0.83
	1	1.00	0.88

7.1 Recall of Invalid Class

We mainly focused on the recall of invalid class as we did not want to miss out a single invalid deduction because predicting invalid deductions as invalid is our first priority.

7.2 Precision of Valid Class

We want our prediction to be as precise as possible. We wish that every class that we classify as valid should actually turn out to be valid.

Table 2 shows that the precision and recall metrics are comparatively much better and accurate when LGBM Technique is used for the model.

Some **other metrics** taken into consideration.

7.3 Leakage

The metric tells about the percent or value of the loss that may occur while evaluating the deductions for the company, it should be minimal or negligible(i.e. ≤ 0.001).

7.4 Autoclear

The metric tells the predicted deduction (prediction from our side whether the deduction is valid or invalid) is same as to the actual deduction (true value), so here how larger the value will be that accurate our prediction is (i.e., Actual = 1 and Predicted = 1).

Table 3 Leakage, autoclear and effortless metrics

Techniques	Leakage	Autoclear	Effortloss
Logistic regression	0.05097327	65.416986	33.348404
Random forest classifier	0.00016458	17.952017	80.813630
Light gradient boosting machine	0.00869187	51.199582	47.566064

7.5 Effortless

The metric tells the predicted deduction (prediction from our side whether the deduction is valid or invalid) is not same as the actual deduction (true value), so here how small the value will be that accurate our prediction is (i.e., Actual = 1 and Predicted = 0).

Let's visualize and compare the results obtained with regards to these metrics:

Table 3 shows us that when the metrics like Leakage, Autoclear and Effortloss are compared with respect to various machine learning algorithms, LGBM technique gives the best results required. According to the study and research work typical business problem with structured tabular data coming from relational database, the LGBM technique is highly preferable because:

1. Light Gradient Boosting Technique (LGBM) is deployed on the basis of decision trees algorithm and splits up the tree leaf-wise which results in much better accuracy as other boosting techniques splits the tree level-wise or depth-wise which have failed to provide good scores.
2. This technique is very fast when it comes to the time consumption for prediction of accurate results in classification problems.
3. Models while making use of LGBM technique consume less memory.
4. And have a good compatibility with larger data sets, i.e., a significant reduction in training time as compared to **XGB** (eXtreme Gradient Boosting).

Hence, out of many classifier techniques used, the LGBM technique has resulted in the best & accurate values.

8 Conclusion

In recent times, machine learning can be helpful in simplifying the process of invoice generation. Here, in this research, various machine learning models were used and implemented for predicting whether a deduction is valid or invalid. It was observed that LGBM model improves on XGBoost. The LightGBM paper uses XGBoost as a baseline and outperforms it in training speed and, therefore, the dataset sizes it can handle. The accuracies are comparable. LightGBM, in some cases, reaches its top accuracy in under a moment and while only reading a fraction of the entire data set.

References

1. Smirnov, J., et al. (2016). Modelling late invoice payment times using survival analysis and random forests techniques. PhD thesis.
2. Tater, T., Dechu, S., Mani, S., & Maurya, C. (2018). Prediction of invoice payment status in account payable business process. In *International Conference on Service-Oriented Computing* (pp. 165–180). Springer.
3. Younes, B. (2013). *A framework for invoice management in construction.* PhD thesis, University of Alberta.
4. Zeng, S., Melville, P., Lang, C.A., Boier-Martin, I., & Murphy, C. (2008). 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.
5. Dirick, L., Claeskens, G., & Baesens, B. (2017). Time to default in credit scoring using survival analysis: a benchmark study. *Journal of the Operational Research Society, 68*(6), 652–665.
6. Lee, E. T., & Wang, J. W. (2013). *Statistical methods for survival data analysis* (4th ed.). Wiley Publishing.
7. Mishra, S., Panda, A., & Tripathy, H. K. (2018). Implementation of re-sampling technique to handle skewed data in tumor prediction. *Journal of Advanced Research in Dynamical and Control Systems, 10*(14), 526–530
8. Sahoo, S., Mishra, S., Mohapatra, S. K., & Mishra, B. K. (2016). Clustering deviation analysis on breast cancer using linear vector quantization technique. *International Journal of Control Theory and Applications, 9*(23), 311–322.