SENSITIVE ATTRIBUTE IDENTIFICATION AND PROTECTION

Submission Date:

File name:

Word Count:

Character Count:

**SENSITIVE ATTRIBUTE IDENTIFICATION AND PROTECTION**

*Report submitted to the SASTRA Deemed to*

*be University as the requirement for the*

*course*

RS300: **RESEARCH CREDIT**

*Submitted by*

**Lakshmi Sri Lasya Tatiraju**

**(Reg No: 125156063, B. Tech CSE AIDS)**

****

**SCHOOL OF COMPUTING**

**THANJAVUR, TAMIL NADU, INDIA – 613 401**

****

**SCHOOL OF COMPUTING**

**THANJAVUR – 613 401**

**Bonafide Certificate**

This is to certify that the report titled “**Sensitive attribute identification and protection**” submitted as a requirement for the course, RS300: **Research Credit** for B.Tech. is a bonafide record of the work done by **Ms. Lakshmi Sri Lasya T (Reg.No.125156063,** **B.Tech.Computer Science & Engineering (Artificial Intelligence & Data Science** during the academic year 2023-2024 , in the School Of Computing, under my supervision

**Signature of Project Supervisor** **:**

**Name with Affiliation** **:** Dr. M. Sumathi

**Date**  **:**

Mini Project Viva-Voce held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Examiner 1 Examiner 2**

**ACKNOWLEDGEMENTS**

We would like to thank our Honorable Chancellor **Prof. R. Sethuraman** for providing us with an opportunity and the necessary infrastructure for carrying out this project as a part of

our curriculum.

We would like to thank our Honorable Vice-Chancellor **Dr. S. Vaidhyasubramaniam** and **Dr. S. Swaminathan**, Dean, Planning & Development, for the encouragement and strategic support at every step of our college life.

We extend our sincere thanks to **Dr. R. Chandramouli,** Registrar, SASTRA Deemed to be University for providing the opportunity to pursue this project.

We extend our heartfelt thanks **to Dr. V. S. Shankar Sriram**, Dean, School of Computing, **Dr. R. Muthaiah**, Associate Dean, Research, **Dr. Ramkumar**, Associate Dean, Academics, **Dr. D. Manivannan**, Associate Dean, Infrastructure, **Dr. R. Algeswaran**, Associate Dean, Students Welfare

Our guide **Dr. M. Sumathi**, **AP-III**, **School of Computing** was the driving force

behind this whole idea from the start. Her deep insight in the field and invaluable suggestions helped us in make progress throughout our project work. We also thank the project review panel members for their valuable comments and insights which made this project better.

We would like to extend our gratitude to all the teaching and non-teaching faculties of the School of Computing who have either directly or indirectly helped us in the completion of the project.

We gratefully acknowledge all the contributions and encouragement from my family and friends resulting in the successful completion of this project. We thank you all for providing me/us an opportunity to showcase our skills through the project.

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure** | **Title** | **Page No** |
| 1.1 | System Architecture | 2 |
| 4.1 | 80th percentile in ‘% Invoices not paid within agreed terms’ | 14 |
| 4.2 | Correlation Analysis | 14 |
| 4.3 | Pie chart of high and low risk before sampling | 15 |
| 4.4 | Pie chart of high and low risk after sampling | 15 |
| 4.5 | ROC curves for all the models | 16 |
| 4.6 | Model comparisons | 16 |
| 4.7 | Feature Importance by XGBOOST in-built function | 17 |
| 4.8 | Feature importance by information value method | 17 |
| 4.9 | Partial dependence plot for ‘average time to pay’ | 18 |
| 4.10 | Partial dependence plot for ‘supply-chain financing offered’ | 18 |
| 4.11 | Partial dependence plot for ‘Payment terms have changed’ | 19 |
| 4.12 | Partial dependence plot for ‘Charges have been made for remaining on supplier list’ | 19 |
| 4.13 | Partial dependence plot for ‘% Invoices paid between 31 and 60 days’ | 20 |
| 4.14 | 10-fold cross validation results on XGBOOST | 20 |
| 4.15 | Results of sampling techniques | 20 |
| 4.16 | Results of various risk proportions | 21 |
| 4.17 | GUI Interface | 22 |
| 4.18 | GUI interface with input and prediction | 22 |
| 4.19 | CNN output with F1 score | 24 |

**ABBREVIATIONS**

**CNN** Convolutional Neural Network

**LSTM** Long Short-Term Memory

**NLP** Natural Language Processing

**BERT** Bidirectional Encoder Representations from Transformers

**RNN** Recurrent Neural Network

**ABSTRACT**

Patient information, such as personal details, details of diseases, and details of treatments, is extremely sensitive in the medical industry. Protecting sensitive information is therefore a fundamental and top need in the medical industry. Typically, entire attributes are safeguarded and kept in external storage locations. The entire attribute protection slows down the processing of the data and decreases its usability for authorized users. Applying protection to a sensitive attribute instead of an entire attribute increases usability for authorized users and protects sensitive attributes. This work aims for the identification of sensitive attributes to achieve this trade-off. Usually, domain experts or machine learning algorithms are used to identify sensitive attributes. These methods’ accuracy is reliant on training data and expert knowledge. In this work, we presented a fuzzy classification-based sensitive attribute identification technique. The attribute values are normalized during the fuzzy classification technique by dividing each value by the maximum value of each relevant attribute. We then developed a brand-new characteristic to classify values into three categories: sensitive, likely sensitive, or not sensitive. The majority category for each attribute decides the sensitivity status. We also implemented CNN, LSTM algorithms to have a comparative analysis. The proposed method is compared with a random forest machine learning classification algorithm. The outcome showed that the suggested fuzzy classification outperformed a random forest with a higher deviation.

**Keywords:** Sensitive Attributes, Medical dataset, Fuzzy classification, Random Forest.

**Table of Contents**

**Title Page No.**

Acknowledgements i

List of Figures ii

Abbreviations iii

Abstract iv

1.Summary of the Base Paper 1

2.Merits and Demerits of the Paper 5

3.Source Code 6

4.Snapshots 16

5.Conclusion and Future Plans 25

6.References 27

7.Appendix-Base Paper 28

**CHAPTER 1**

**SUMMARY OF THE BASE PAPER**

* 1. **INTRODUCTION**

On average the world generates 2.5 quintillion bytes per day. There is a wide range of domains that generate sensitive data. In this highly growing data world, there is a high necessity of identifying sensitive data and also protecting the data as well. However, storing all the data that is generated is a tough task to do. Hence there has to be a filter to identify only sensitive data to reduce the slowing down of systems and also to enhance usability. The techniques used to generate sensitive data have to concentrate on identifying the sensitivity level for each attribute so that the threshold to cluster sensitive attributes will be decided by the domain experts of the respective domains.

* 1. **RELATED WORK**

In this section, the authors have displayed a summary of the literature review. [1] Addresses the concern regarding sensitive data extraction by implementing a fuzzy logic system. Quasi-dependent and quasi-independent attributes are identified and then rules are built over membership functions. Fuzzy logic can handle subjective data security assessments. [3] Is based on deep neural networks and it specifically works on financial data. The datasets are manually tagged and sent to models like CNN and LSTM to train the model. This model predictions are of two types; one is entity detection and the other is column-wise entity detection on tabular data. Accuracy is considered as a measure to compare the models.

[2] Implements language representation models like BERT that employ bidirectional Transformer encoders for deep contextual representation learning. Tokenization, embedding and classification of sequences of text are done internally by BERT. It becomes task-specific by adding an output layer. In contrast, authors’ show the concern towards sensitive data extraction by implementing deep learning models like CNN, and LSTM and machine learning models like random forest, decision tree and fuzzy logic on generalized data and the models are compared based on accuracy and also on a number of sensitive attributes identified. This approach identifies sensitive data from both structured and unstructured formats.

* 1. **PROBLEM STATEMENT**

The entire attribute protection slows down the processing of the data and decreases its usability for authorized users. Applying protection to a sensitive attribute instead of an entire attribute increases usability for authorized users and protects sensitive attributes.

* 1. **OBJECTIVES**

To classify sensitive attributes from a structured and semi-structured data by using fuzzy-rules and natural language processing techniques and also ML and DL techniques. Applying protection to sensitive attributes by using attribute-based encryption**.**

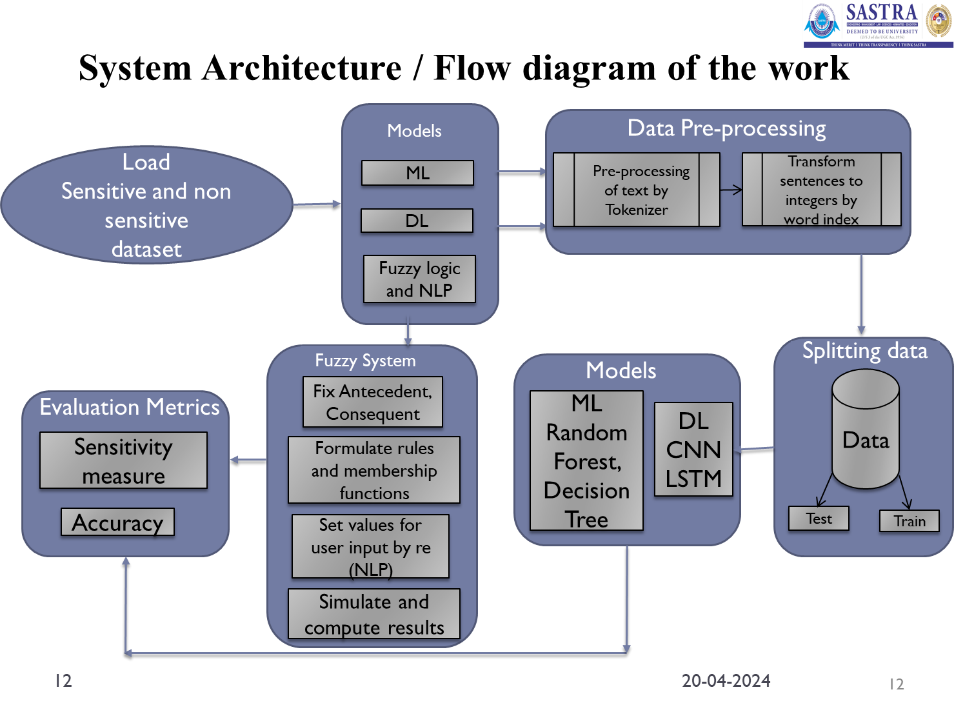
* 1. **PROPOSED SOLUTION AND SYSTEM ARCHITECTURE**

This work addresses the issue by developing models from deep learning, machine learning and fuzzy logic implementation. In this work, there is a comparative analysis for sensitive data extraction using deep learning models CNN, LSTM and machine learning algorithms like random forest, decision tree, and transformer encoders like BERT. Data-preprocessing steps are followed the same for all the models to maintain a justified comparative analysis. Data is in the form JSON format and any other formats are converted to JSON format and machine and deep learning models are trained and tested over them. The authors identified the system architecture as shown in Figure 1.1.

* 1. **HARDWARE AND SOFTWARE REQUIREMENTS**
* **Hardware:**
  + 4GB RAM
  + 256GB external memory
* **Software:**
  + Visual Studio (VS) Code
  + Python

* 1. **METHODOLOGY AND IMPLEMENTATION**

The sensitive and non-sensitive JSON files are taken separately and are preprocessed to input the models for training. Data in JSON files is of the form dictionaries hence they are loaded and converted in the form of data frames. The dictionary has two keys one is 'is\_sensitive' which tells about the sensitivity of data and the other is 'data' which consists of data. Now the is\_sensitive will be the target label for the data in the data frame. Both the classes of data are down sampled to promote balance between the two classes and then they are concatenated and then split into training and validation sentences. Now after tokenization and conversion to a word index. The data is fed to models like CNN, LSTM, Random Forest, Decision Tree, Bert and also a fuzzy system to identify accuracy and also to identify sensitive attributes from unseen data. for the fuzzy system, NLP techniques like regular expression matching [5] are implemented in the first place to identify a few sensitive attributes like 'username', 'password', 'phone number', and 'email-id' and then antecedent and consequent are fixed. Later these inputs are fed to membership functions and then rules are formulated. After simulation results are obtained for a given input. After all the evaluation sensitive attributes are protected by implementing Fernet symmetric encryption method.



**Fig 1.1 System Architecture**

The authors have proposed to develop a model for sensitive attribute identification fuzzy logic, in addition to that in this work the fundamental models taken reference from various sources and are modified and directed to be implemented for text classification purposes. The models are evaluated based on accuracy and the count of identification of sensitive attributes over unseen data.

After the data pre-processing like tokenization and word-index generation, sequences are generated and are padded to input the models. The data is balanced by down sampling and then 20% of the data is considered for testing and the remaining 80% is used for training.

The models are implemented on generalized data and also over healthcare domain data. The architecture used is CNN is 1D CNN with an Embedding layer at input followed by Batch Normalization, Convolution, Dropout, Global Average Pooling, and Dense Layers. RNN is implemented using LSTM units. It starts with an embedding layer to convert inputs to dense vectors, and then LSTM captures sequential information and the final dense layer classifies the input text. In the BERT model, forward method defines the forward pass of the model: it takes in sentence IDs and an attention mask, passes them through the BERT model, and then through two linear layers with a ReLU activation function and dropout in between. The output of the final linear layer is passed through a softmax function to produce the final classification probabilities. Decision trees and random forest classifiers are implemented over pre-processed data. A fuzzy logic system is built over input data and rules are formulated. Later by the values of membership functions, they are evaluated and sensitivity is defined.

**CHAPTER 2**

**MERITS AND DEMERITS OF THE BASE PAPER**

**2.1 MERITS**

The proposed work not only implements all the above-mentioned elements but also includes several other aspects as well. The following are the merits of the proposed work.

1. **Data availability** is a crucial problem in dealing with sensitive data related works. This problem is well tackled here by finding a large, generalized sensitive data in a convenient JSON format which is easier for the models to train and test as well.
2. **Implementation of CNN** ensures understanding of patterns available and also there is no need of special feature extraction. **LSTM** identifies the long-term dependencies in the sequential text hence there won’t be any error when the text has long term dependencies.
3. **BERT's** deep bidirectional understanding of context in text data can lead to highly accurate text classification, even with complex or nuanced categories.
4. **Fuzzy** long doesn’t require training and testing hence it can used when there is lack of data.

**2.2 DEMERITS**

The proposed work consists of multiple areas of strength when compared to the existing solutions. However, some areas could be improved as follows:

1. Deep models are prone to over fitting and the training of the models
2. Fuzzy logic rules are subjective and there has be a manual check over the rules for including all possible cases.

**CHAPTER 3**

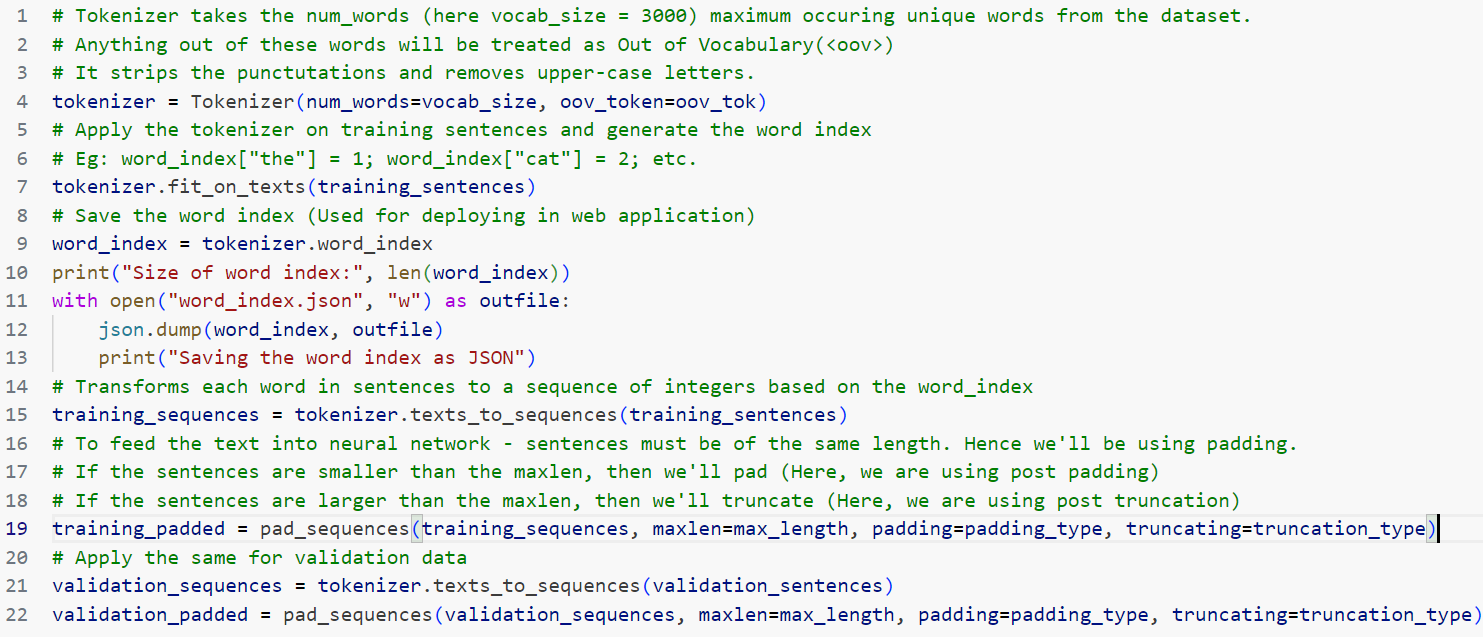
**SOURCE CODE**

**3.1 Data Preprocessing**

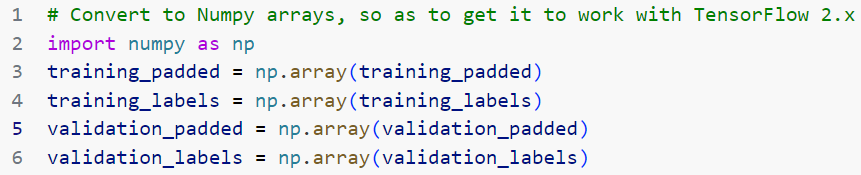
**Spliting data after formulating sensitive and non-sensitive data frames and splitting:**

****

**Tokenization, word-indexing, formulating sequences:**

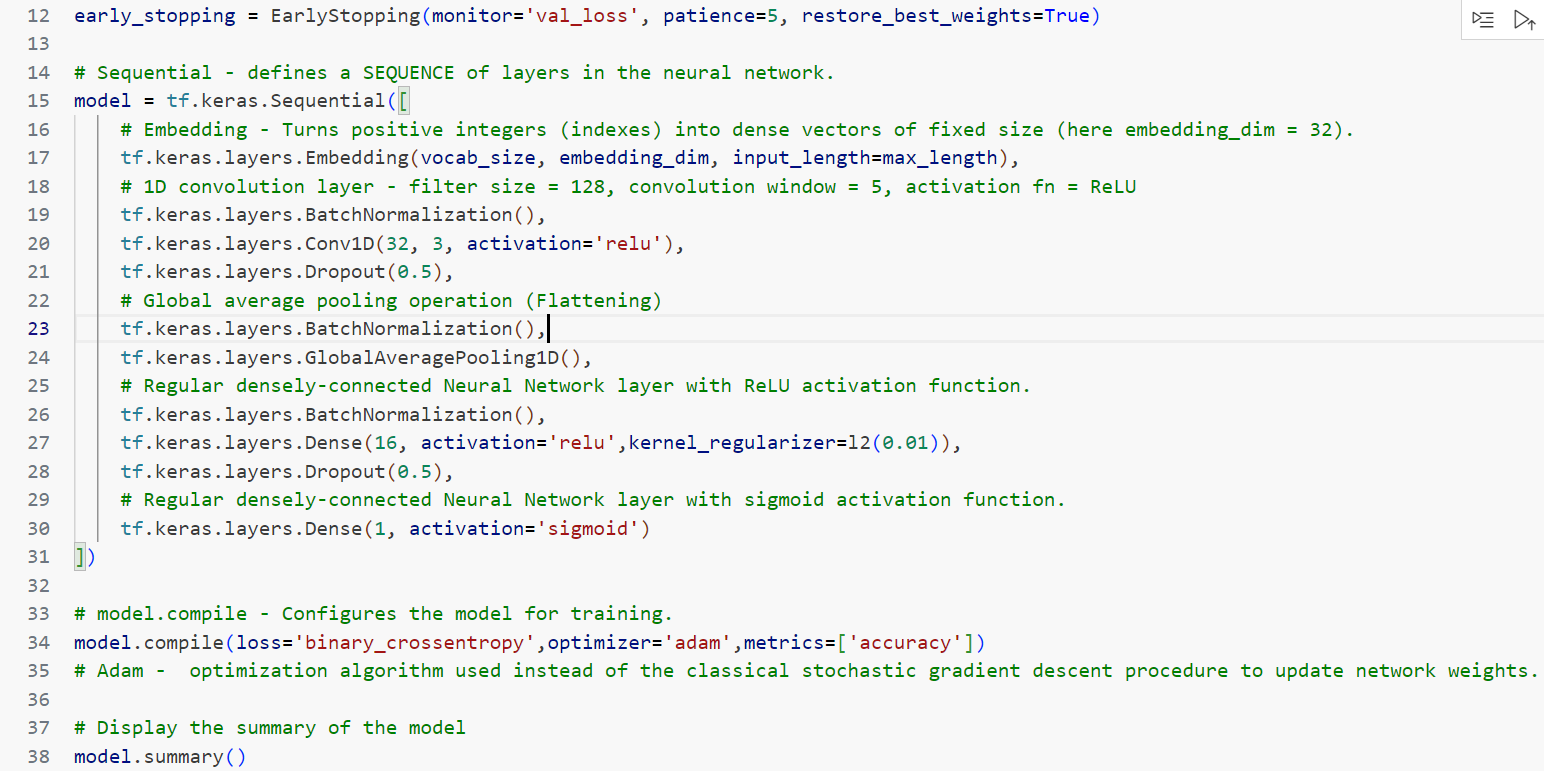
****

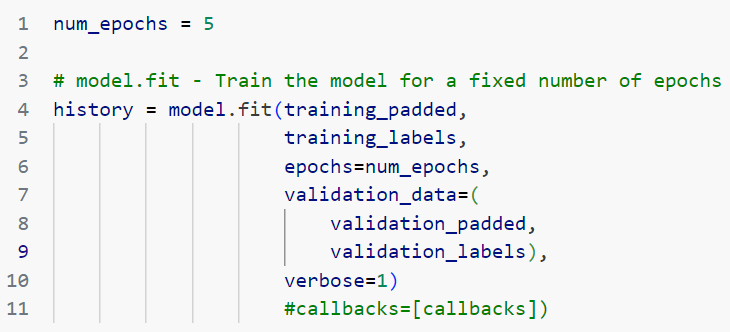
**Converting to NumPy array:**

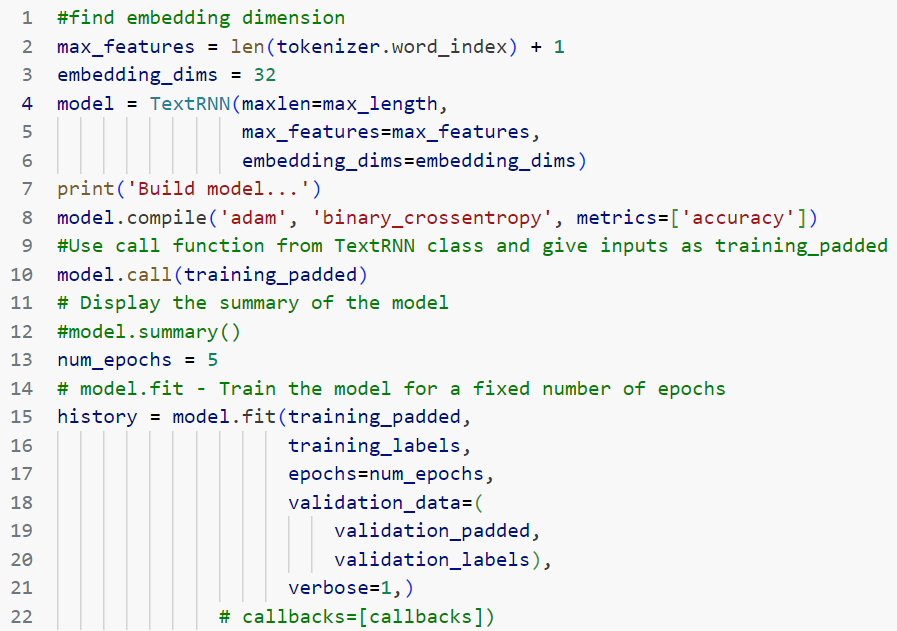
****

**3.2 Models Implementation**

**CNN model:**

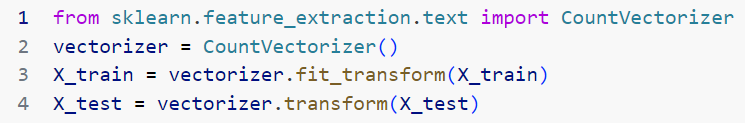
****

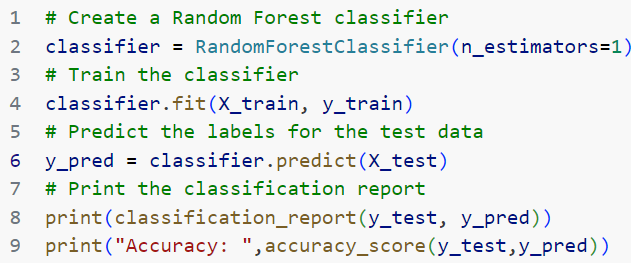
****

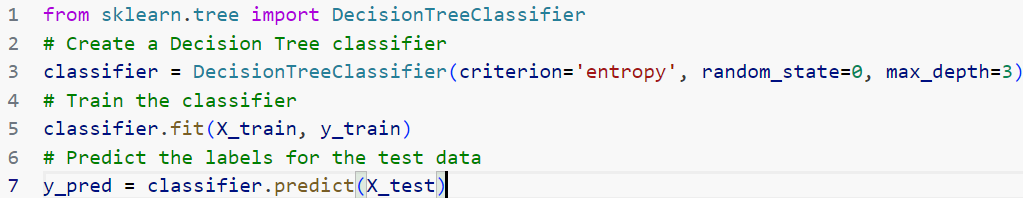
**LSTM model:**

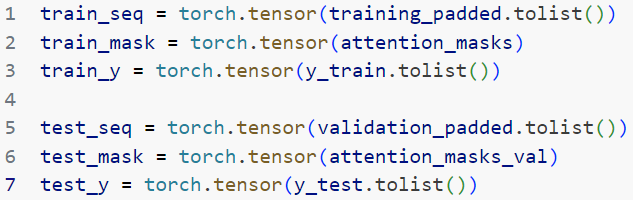
**Random Forest and Decision Tree:**

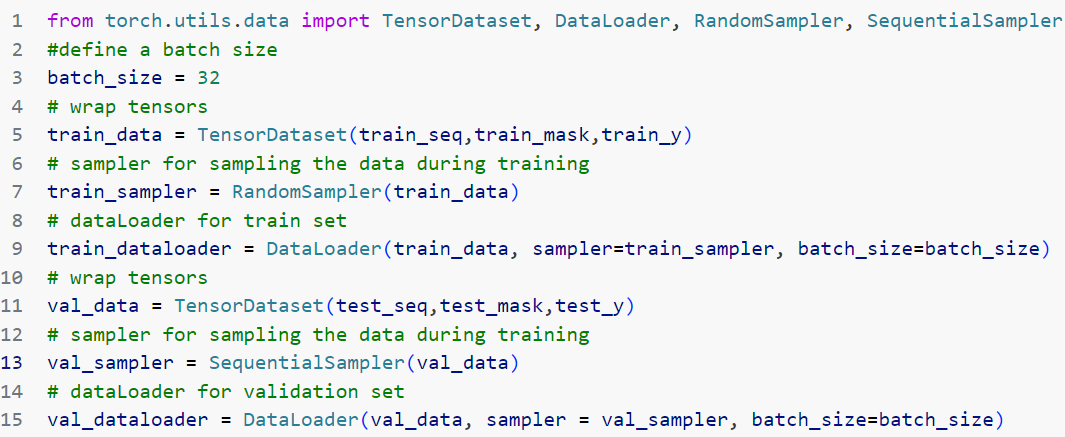
****

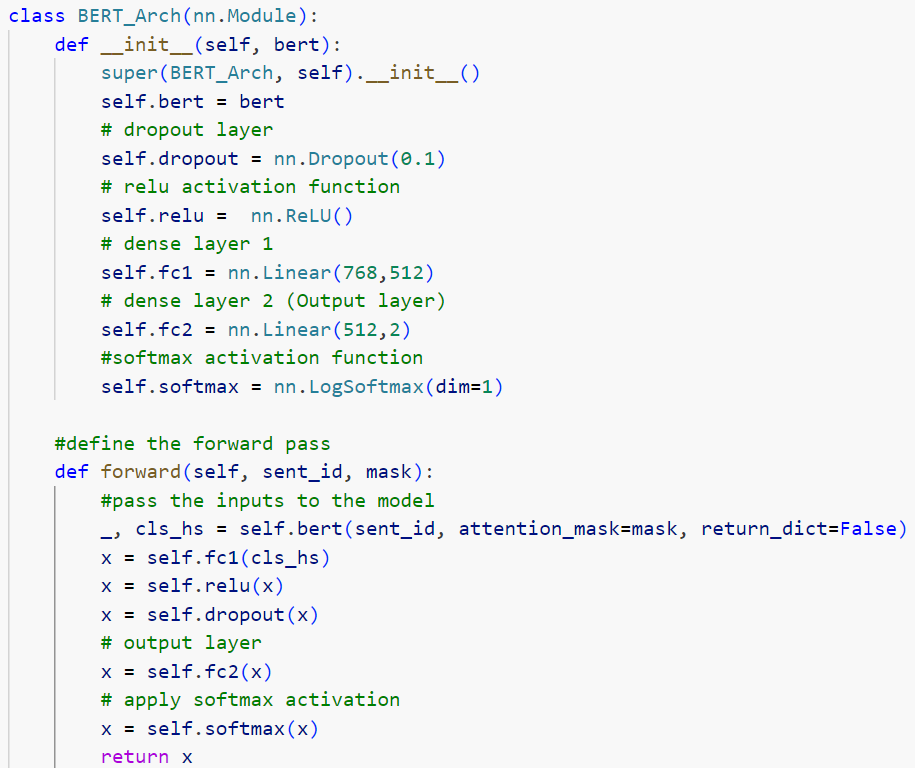
****

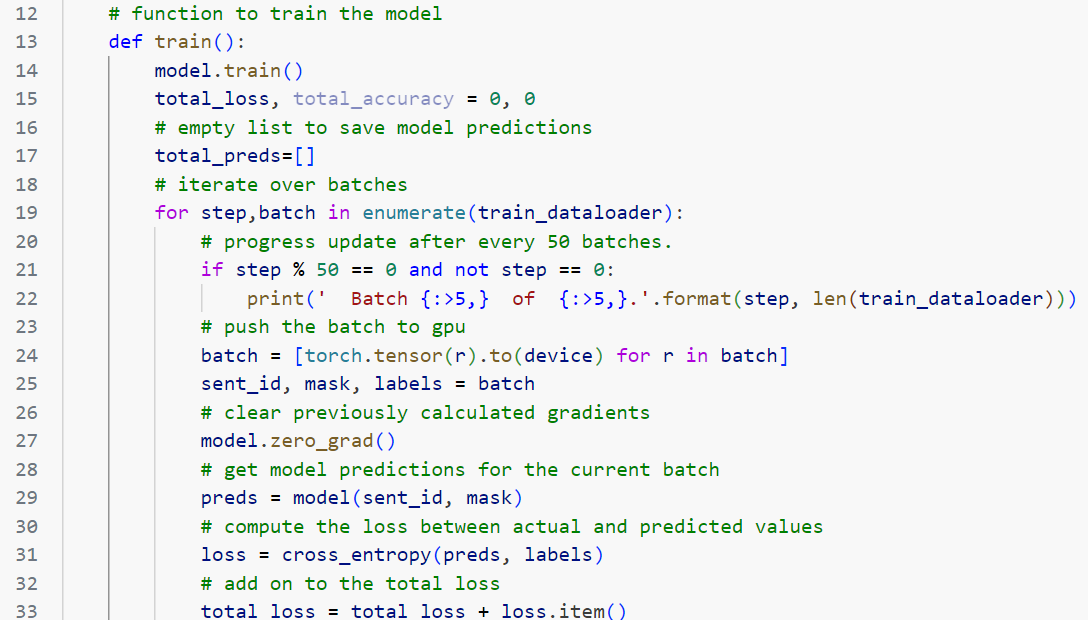
****

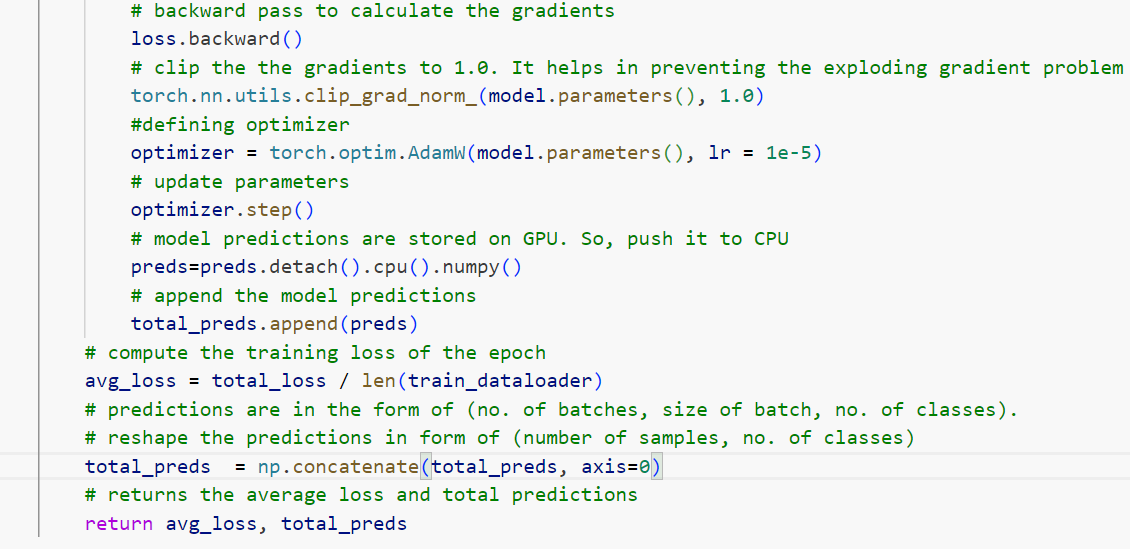
**BERT Model:**

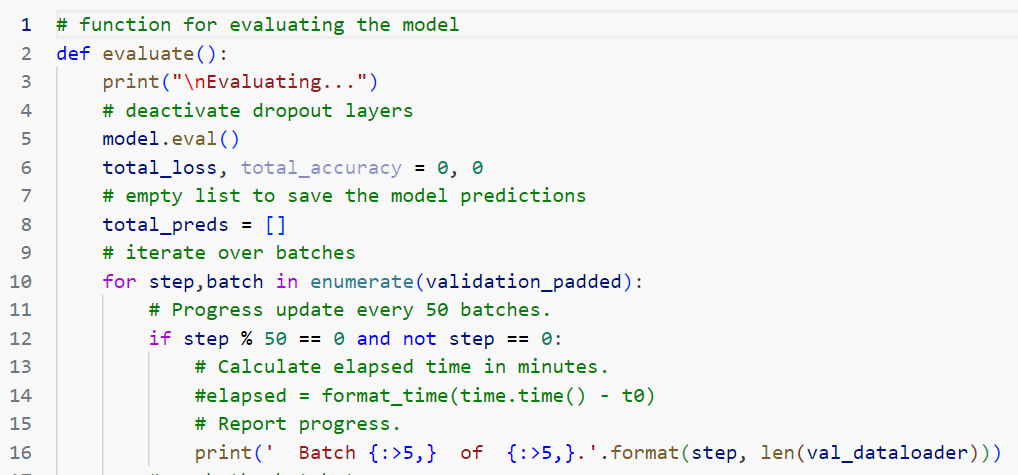
****

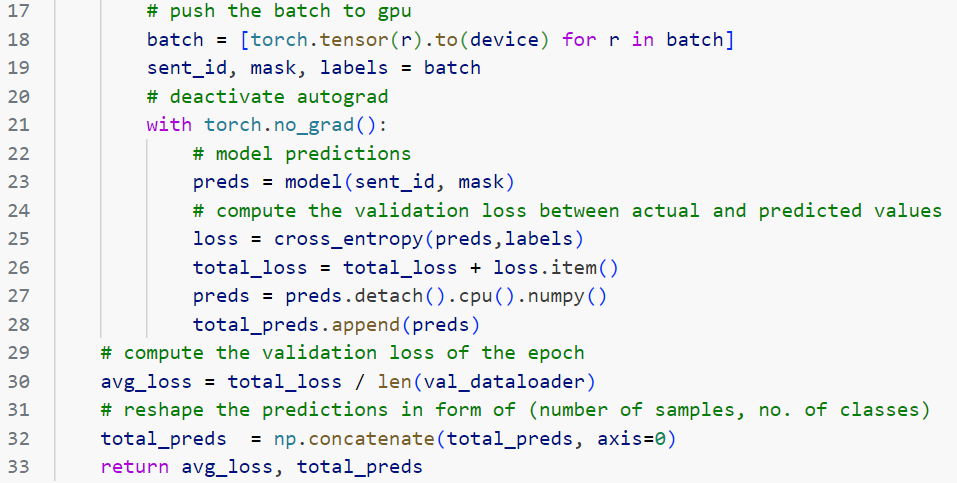
****

****

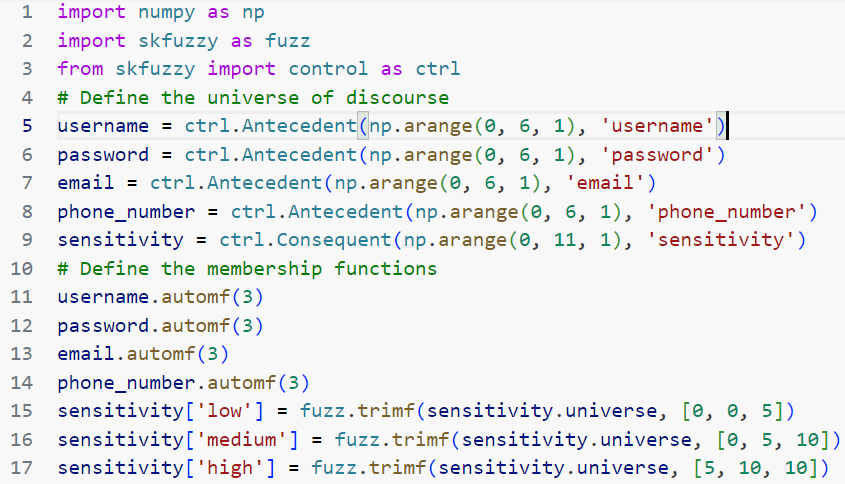
****

****

****

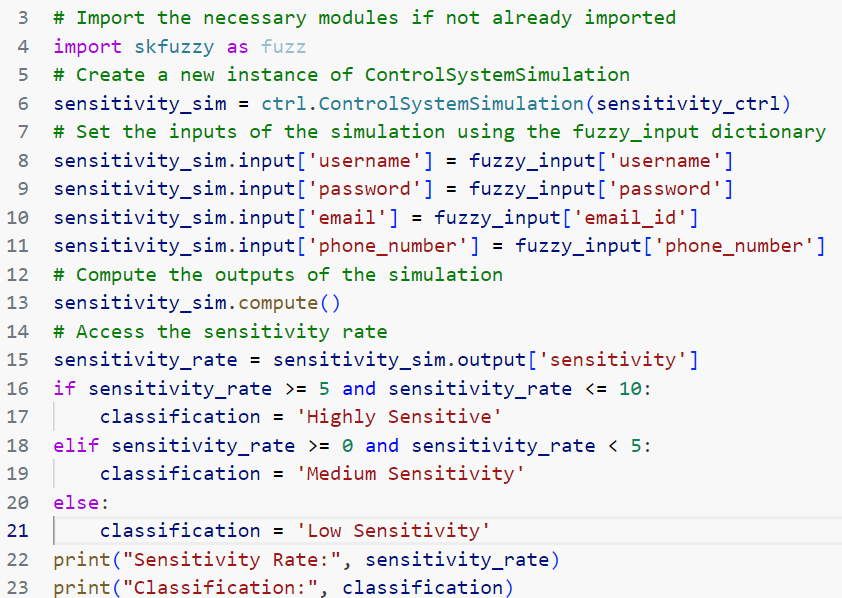
****

**Fuzzy system:**

****

****

****

****

**CHAPTER 4**

**OUTPUT SNAPSHOTS**

**Bar plot for '% Invoices not paid within agreed terms' and 80th percentile in figure 4.1**

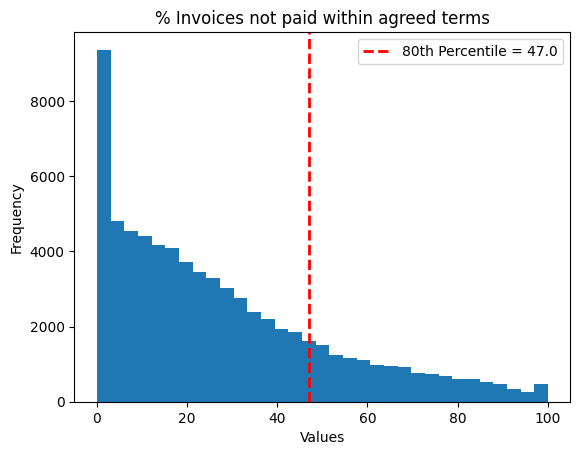


Fig.4.1 80th percentile in ‘% Invoices not paid within agreed terms’

The figure explains about the 80th percentile in ‘% Invoices not paid within agreed terms’ to identify the proportion of high risk and low risk samples in target variable. Initial consideration was 20 % high risk and 80% low risk, later there is an analysis made for different proportions.

**Correlation Analysis in figure 4.2**

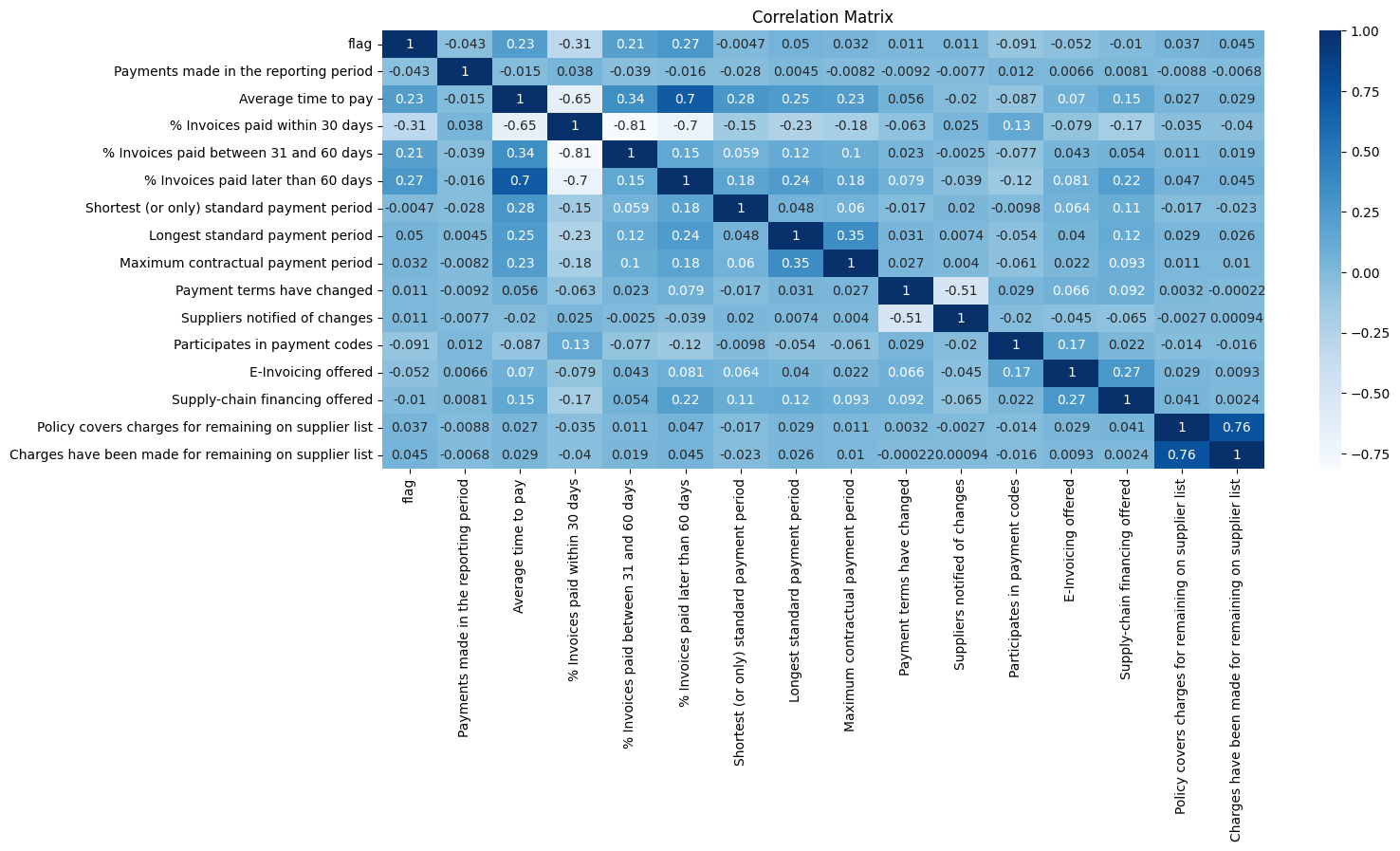
****

Fig.4.2 Correlation Analysis

The correlation matrix explains about correlation between features. Highly correlated features are analyzed and are removed to avoid unnecessary computations. We observed that

1. % Invoices paid within 30 days is highly correlated with % Invoices paid between 31 and 60 days (correlation: -0.81)
2. % Invoices paid between 31 and 60 days is highly correlated with % Invoices paid within 30 days (correlation: -0.81)
3. Policy covers charges for remaining on supplier list is highly correlated with Charges have been made for remaining on supplier list (correlation: 0.76)
4. Charges have been made for remaining on supplier list is highly correlated with Policy covers charges for remaining on supplier list (correlation: 0.76)

**Pie chart of high and low risk before sampling in figure 4.3**

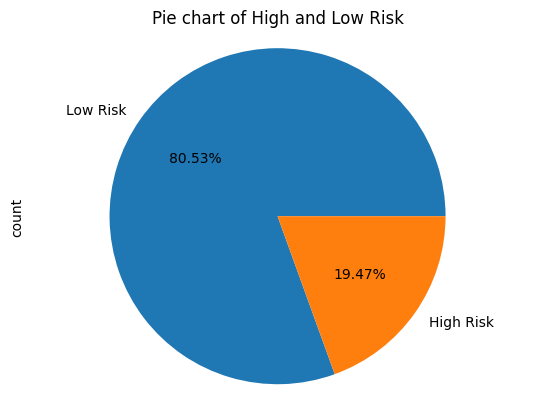
****

Fig.4.3 Pie chart of high and low risk before sampling

There is an analysis performed to identify high risk and low risk samples within the target for identification of imbalance in the data. There are several re-sampling techniques like random over sampling, random under sampling and SMOTE to balance the data.

**Pie chart of high and low risk after sampling in figure 4.4**

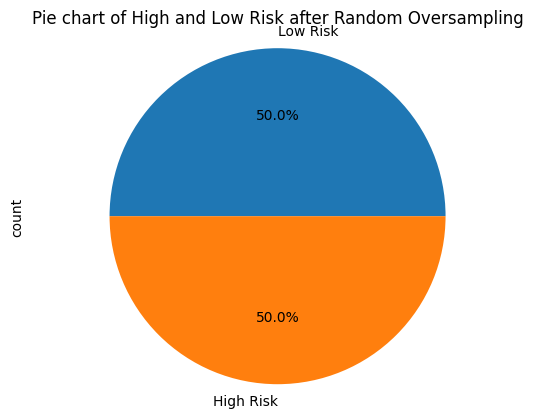


Fig.4.4 Pie chart of high and low risk after sampling

After observing the results of different re-sampling techniques, random over sampling out performs with F1 -score of 54.9 which is highest among random under sampling and smote.

**ROC (Receiver Operating Characteristic Curves) in figure 4.5**

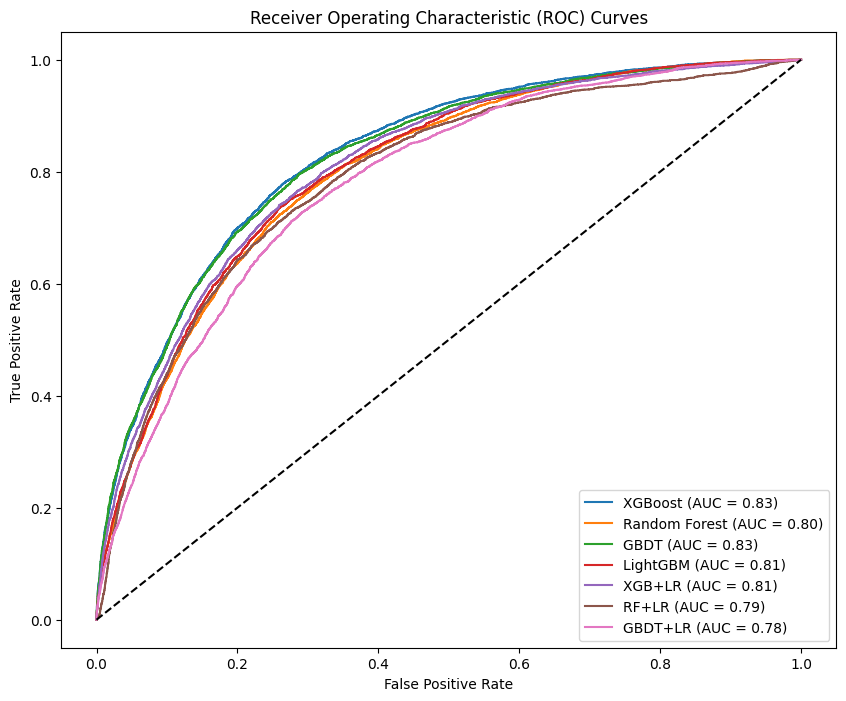
****

Fig.4.5 ROC curves for all the models

By plotting ROC curve, we can identify that XGBOOST has highest AUC of 0.83, followed by GBDT with AUC of 0.83 and Random Forest with AUC of 0.80. We can observe that single models perform well than hybrid models.

**Model comparisons is figure 4.6**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Model** | **Accuracy** | **F1 Score** | **AUC** |
| **1** | **Random Forest** | **0.700144** | **0.509270** | **0.802509** |
| **2** | **GBDT** | **0.774579** | **0.552037** | **0.826370** |
| **3** | **LightGBM** | **0.717474** | **0.517679** | **0.806924** |
| **4** | **XGB+LR** | **0.746640** | **0.532741** | **0.812604** |
| **5** | **RF+LR** | **0.802477** | **0.493552** | **0.792003** |
| **6** | **GBDT+LR** | **0.671528** | **0.486486** | **0.782524** |
| **7** | **XGBOOST** | **0.762955** | **0.551432** | **0.830329** |

Fig.4.6 Model comparisons

**Feature Importance using XGBOOST in-built function in figure 4.7**

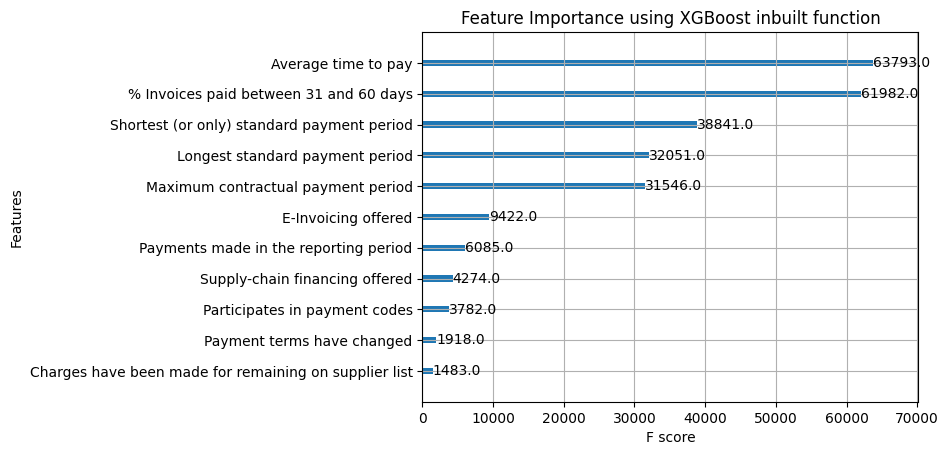
****

Fig.4.7 Feature Importance by XGBOOST in-built function

This is feature importance plot made using inbuilt XGBOOST function to make understand what features are essentially prevailing risk in supply chain. This is an XAI method brought here to make model interpretable and explainable. Here it tells about target variable dependence on each feature.

**Feature importance by IV in figure 4.8**

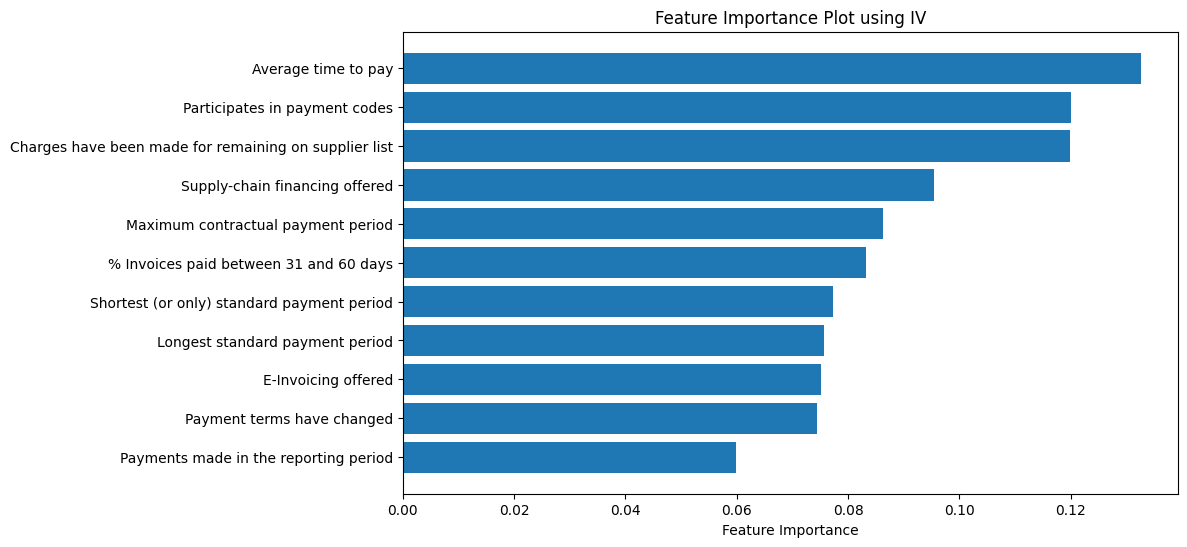
****

Fig.4.8 Feature importance by information value method

This is also a partial dependence plot which interprets the dependence of target on each feature but here it is made using information value method in which features importance are obtained by XGBOOST in-built function and the values are sorted for knowing the highly important feature.

**Partial dependence plot for ‘average time to pay’ in figure 4.9**

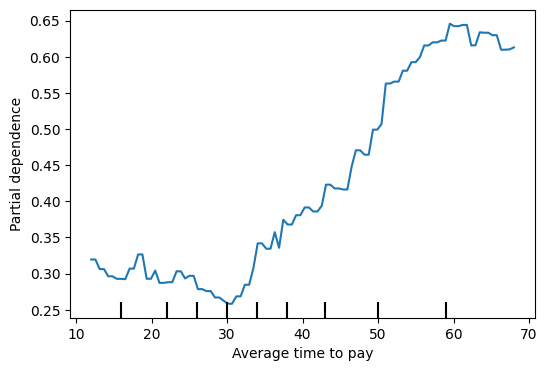
****

Fig.4.9 Partial dependence plot for ‘average time to pay’

This is the individual trend of dependence of output on the feature ‘average time of time’ which is identified as most important by partial dependence plot.

**Partial dependence plot for ‘supply-chain financing offered’ in figure 4.10**

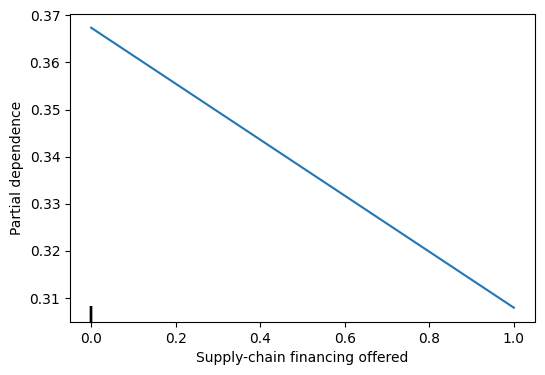


Fig.4.10 Partial dependence plot for ‘supply-chain financing offered’

There is a negative dependence of target on the feature ‘supply-chain financing offered’ yet this feature is important in risk assessment.

**Partial dependence plot for ‘payment terms have changed’ in figure 4.11**

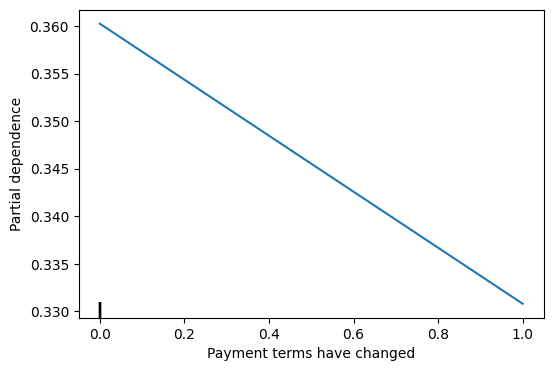
****

Fig.4.11 Partial dependence plot for ‘Payment terms have changed’

There is a negative dependence of target on the feature ‘payment terms have changed’.

**Partial dependence plot for ‘charges have been made for remaining on supplier list’ in figure 4.12**

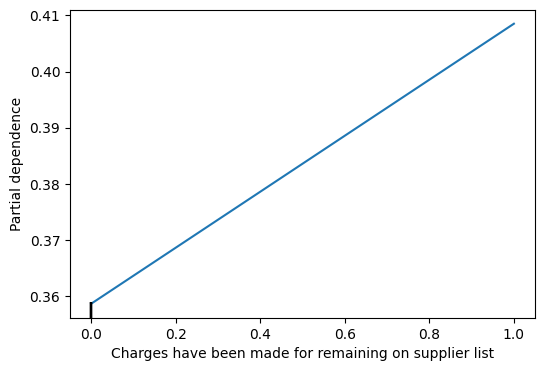


Fig.4.12 Partial dependence plot for ‘Charges have been made for remaining on supplier list’

There is a positive dependence of target on the feature ‘Charges have been made for remaining on supplier list’.

**Partial dependence plot for ‘% Invoices paid between 31 and 60 days’ in figure 4.13**

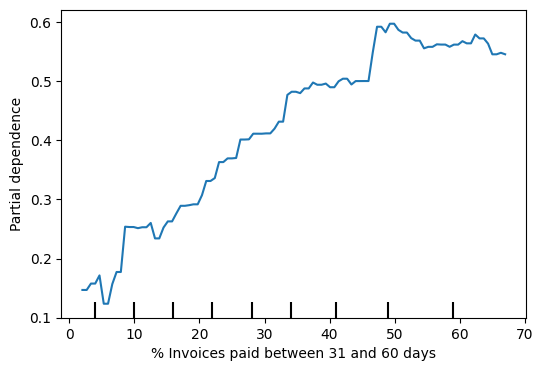


Fig.4.13 Partial dependence plot for ‘% Invoices paid between 31 and 60 days’

There is an increasing trend observed in identifying dependence of target on ‘% Invoices paid between 31 and 60 days’

**10-fold cross-validation results on XGBOOST in figure 4.14**

|  |  |
| --- | --- |
| **Average Accuracy** | **0.8129338704884527** |
| **Average F1 Score** | **0.8217140268262384** |
| **Average ROC AUC Score** | **0.8129096994521376** |

Fig.4.14 10-fold cross validation results on XGBOOST

10-fold cross-validation gives a good F1-score of 82.17 which is 38.9% higher than train-test split of data for XGBOOST.

**Results of sampling techniques in figure 4.15**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Model** | **Accuracy** | **F1 Score** | **AUC** |
| **0** | **Random Oversampling** | **0.774966** | **0.549779** | **0.814912** |
| **1** | **Random Undersampling** | **0.706352** | **0.524085** | **0.802064** |
| **2** | **SMOTE** | **0.797724** | **0.516050** | **0.805171** |

Fig.4.15 Results of sampling techniques

Random over sampling shows a good increment in F1-score with 4.9% over random under sampling and an increment of 6.5% over SMOTE procedure.

**Results of various risk proportion in figure 4.16**

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk Proportion** | **Accuracy** | **F1 Score** | **AUC** |
| **10%** | **0.860599** | **0.430976** | **0.831349** |
| **20%** | **0.801072** | **0.553533** | **0.825286** |
| **30%** | **0.759830** | **0.620794** | **0.814654** |
| **40%** | **0.740995** | **0.687769** | **0.812988** |
| **50%** | **0.750069** | **0.753123** | **0.824772** |

Fig.4.16 Results of various risk proportions

By observing several proportions of high risk and low risk samples, 20% gives stable results and is also relevant with data.

**GUI interface in figure 4.17**

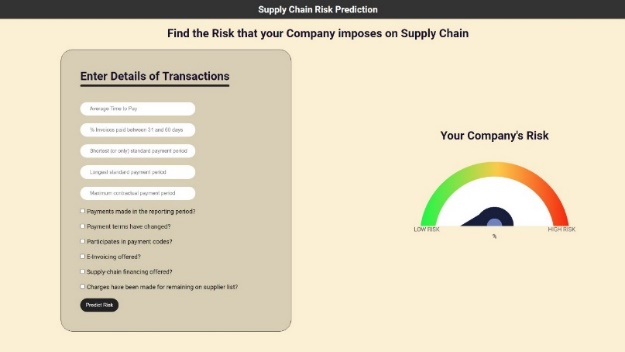


Fig 4.17 GUI Interface

This is the GUI interface made for the user convenience. The web page consists of features essential for evaluating. Few features like ‘average time to pay’, ‘% invoices between 31 and 60 days’, ‘shortest or only standard payment period’, ‘longest standard payment period’, ‘maximum contractual payment period’ are taken numerical input which other binary features are given as options to select.

**GUI interface with input and prediction in figure 4.18**

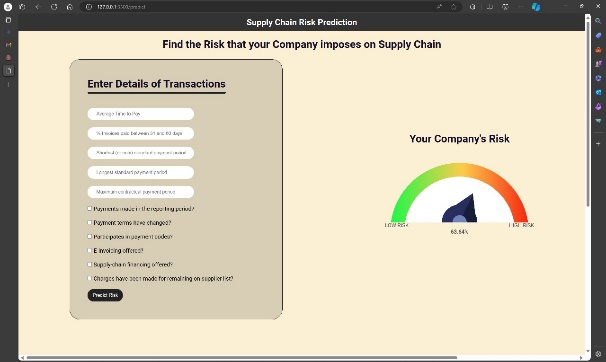
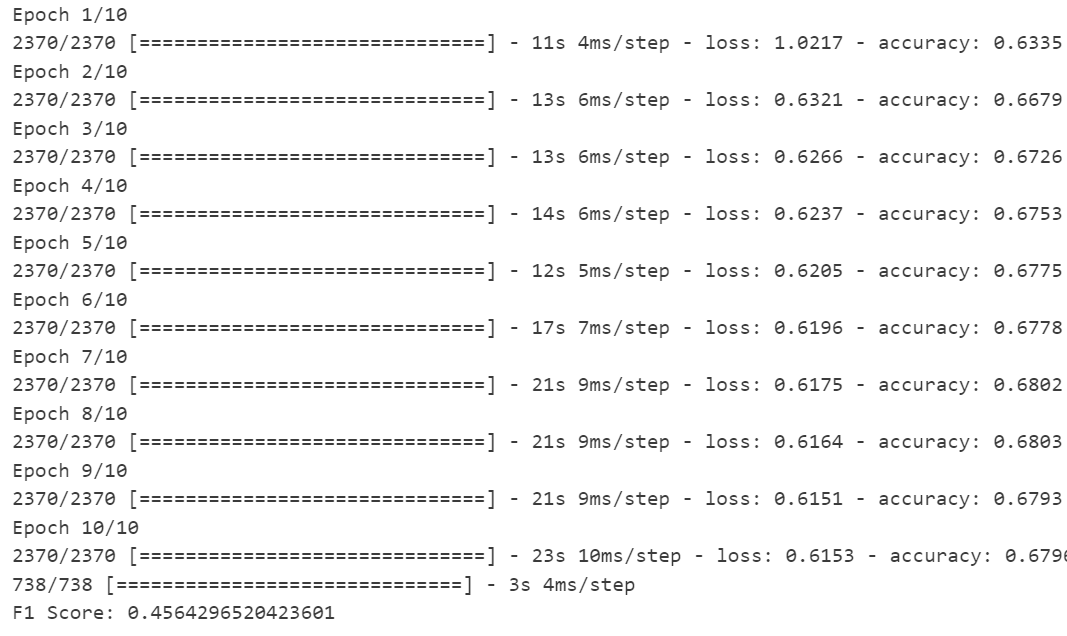


Fig 4.18 GUI Interface with input and prediction

This is the final web page with the input and prediction. The risk prediction obtained here is 68.54% which is also indicated with mild red which resembles the impact of risk.

**CNN output with epochs count is in figure 4.19**

****

The CNN model implemented with regularization is obtained with the F1 score of 45.6 and there were 10 epochs.

**CHAPTER 5**

**CONCLUSIONS AND FUTURE PLANS**

**5.1 CONCLUSIONS:**

Today supply chain finances face several challenges, especially concerning delayed payments by customers. Understanding customer transaction behavior and collaborating that knowledge with machine learning and deep learning models can enhance the chances of predicting risk in the supply chain. Displaying the results with proper interpretations and explanations can secure the company from risk and also develop trust in the reason for the risk that was predicted. When tree-based machine learning models are used, the results are quite explanatory as the decision-making is transparent to understand. This gives companies a good precaution for the risk that is going to prevail and gives a hint for the method to avoid risks by making an idea over features that are contributing to the risk factor.

The proposed supply chain risk prediction system puts forward risk prediction through a website that takes relevant features to evaluate the risk and display the amount of risk using a risk meter for user convenience and feasibility. Features like ‘average time to pay’,’% Invoices paid between 31 to 60 days’, ’shortest payment method’ and ‘longest payment method’ are asked to enter numerical values and there is a check box given for user convenience to select features like ‘Payments made in the reporting period’, ‘Payment terms have changed’,’ Participates in payment codes’, ‘E-Invoicing offered’, ‘Supply-chain financing offered’, ‘Charges have been made for remaining on supplier list’ is of binary (yes or no). All these features are identified by the feature importance model and XGBOOST outperforms all the machine learning algorithms with an F1-score of 0.55 which is 8.2% higher than Random Forest, and 3.5% higher than XGBOOST + LR. As XGBOOST outperforms, it is improvised with a different pre-processing technique of imputing data with ‘iterative imputer’ which showed 0.59 as the F1-score which is 7.9% higher than the previous XGBOOST model. The approach also includes a CNN model implementation with the same data pre-processing to identify the F1-score. XGBOOST outperforms CNN by an increment of 12% in the F1-score (CNN F1-score: 0.49). Hence the inputs from the website are passed to the XGBOOST algorithm as a test case and the prediction percentage evaluated is displayed numerically and also in the risk meter.

The data considered for analysis is from the UK government and it is dynamic data. Hence an analysis is made and model implementation is done over both the base paper dataset (till 2018) and the updated data set (till 2023). XGBOOST acquired an F1-score of 0.547 and AUC of 0.812 on old data while on the new data set F1-score is about 0.551 and AUC is about 0.830.

In conclusion, the proposed application overcomes the challenges of unclear results and complex methods of implementation for predicting risk in the supply chain. This helps for the smooth balance of the financial aspects of the company. This is a user-interactive system that helps non-technical users to easily evaluate their risk rate.

**5.2 FUTURE PLAN:**

The following methods can be employed in future:

1. Direct the methods towards data independent approach which gains good attention when there is lack of data.

2. Model independent results

**CHAPTER 6**

**REFERENCES**

**[1]** Ivanov, D., & Dolgui, A. (2019). Low-Certainty-Need (LCN) supply chains: a new perspective in managing disruption risks and resilience. International Journal of Production Research, 57(15–16), 5119–5136. https://doi.org/10.1080/00207543.2018.1521025

**[2]** Guan, X., Li, G. & Yin, Z. The implication of time-based payment contract in the decentralized assembly system. *Ann Oper Res* **240**, 641–659 (2016). <https://doi.org/10.1007/s10479-014-1579-5>

**[3**] Zelong Yi, Zhuomin Liang, Tongtong Xie, Fan Li, Financial risk prediction in supply chain finance based on buyer transaction behavior, Decision Support Systems, Volume 170, 2023, 113964, ISSN 0167-9236, <https://doi.org/10.1016/j.dss.2023.113964>.

**[4]** R. Qin, The construction of corporate financial management risk model based on XGBoost algorithm, J. Math. 2022 (2022) 1–8, <https://doi.org/10.1155/2022/2043369>.

**[5]** M. Bahrami, B. Bozkaya, S. Balcisoy, using behavioral analytics to predict customer invoice payment, Big Data 8 (1) (2020) 25–37, https://doi.org/10.1089/ big.2018.0116.

**CHAPTER 7**

**APPENDIX – BASE PAPER**

**Title :** Financial risk prediction in supply chain finance basedon buyer transaction behavior

**Author :** Zelong Yi, Zhuomin Liang, Tongtong Xie, Fan Li

**Year of Publication :** 2023

**Journal Name :** Decision Support Systems 170 (2023) 113964

**Indexed in :**  Science Citation Index Expanded