Credit Default Prediction

Pravin Pawar, Priyesh FNU
pawar.prav@northeastern.edu, priyesh.p@northeastern.edu

Abstract

Credit risk prediction with accuracy in the dynamic real world is always a **crucial task** in the **finance industry**, as incorrect prediction can lead to **loss in billions of dollars** along with **rise in inflation** and **market downfall**. While most of the Machine models make predictions based on application level itself with a large number of insignificant features, we have experimented with **GNN[2] and LLMs based approaches**[1] which take in account similarity between applicants and accumulate its relationships to provide improved feature embedding and provide better performance for Machine Learning Models. We have also experimented with a **Multi-model ensemble**[4] of various advanced classification models such Neural Network, Random Forests, KNN, etc and performance experiments on **Explainable AI**[4] where we are gauging which features have higher weightage or significance for specific predictions. Lastly, we have to handle **class imbalance**[3] **and dimensionality reduction** to have more accurate predictions required for dynamic real work data. Experiments based on above approaches have shown promising results with accuracy surpassing industry benchmarks within required standard of confidence intervals providing more efficient and robust models suitable to incorporate uncertainty of the real world.

Introduction

With constantly changing real-world, credit risk prediction has been a challenging task faced by financial institutes on a daily basis. In 2008, there was a huge financial downfall because credit risk of consumers was not predicted in time leading to increasing flow of credit as well as increasing risk. Resulting in **losses in billions of dollars** as well as collapse of the **World financial market**, it took years for the financial market to recover. Thus, accurate risk prediction is the most significant task for sustainable cash flow in the **global market affecting the lives of people all across the world**. Credit risk is influenced by uncertain factors like unseen events, market fluctuation and changes in regulations making it complex tasks to solve. Due to advancement of technology and computation, advance modelling using rule based programming and machine learning models have been utilised to improve predictions.

However, most of the models used today work on single application level information while prediction. Currently, researchers are focusing on **capturing relationships[2]** between data points as well as providing **explainability[4]** behind specific decisions to make firm decisions. On this project, we are focusing on analysing and enhancing credit risk prediction considering as classification task whether candidate will default in coming 60 days or not by merging information on highly similar applications to improve application profile embedding or feature vectors by implementing and experimenting with advanced machine learning algorithms like GNNs, LLMs, Multi-model ensemble of SVC, Deep Neural Network, etc. Also, We are performing an experiment using a multi-model ensemble of advanced machine models to fit a different distribution of data points using various modelling algorithms. We have also focused on explainability of model prediction by providing details of which features were given

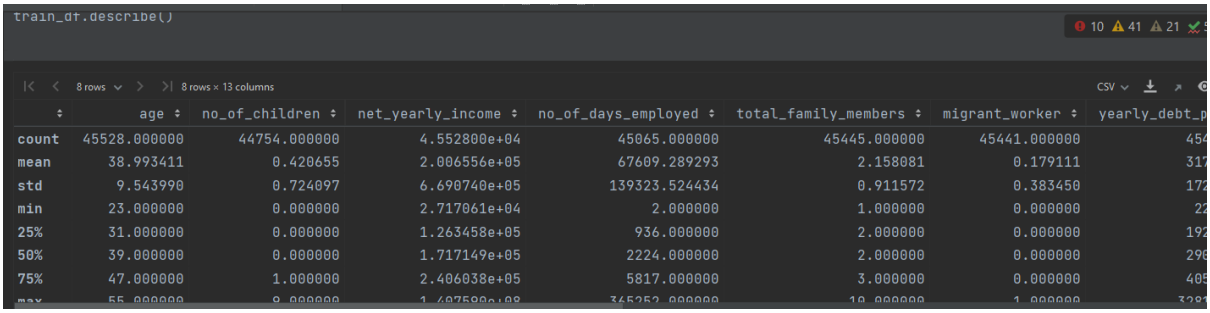
importance while making specific decisions by model helps to identify any biases in decision making[3].

Grid Search cross validation has provided improved accuracy with best fit parameters. Multi model ensemble of standard Algorithms to provide better generalisation on test set as compared to standard algorithms. Accuracy is further improved using Graph neural network taking relationship customers into account. At the end, LLMs based model seems to provide optimised feature representation with highest accuracy among all other models.

Experiment setup

Dataset: American express credit card dataset containing details of real world customers is used for this project. It consists of 45k records of customers with core features like credit score, occupation type, number of children, annual income, employment days, etc.
Exploratory Analysis (Key statistic) of datasets are as follows:

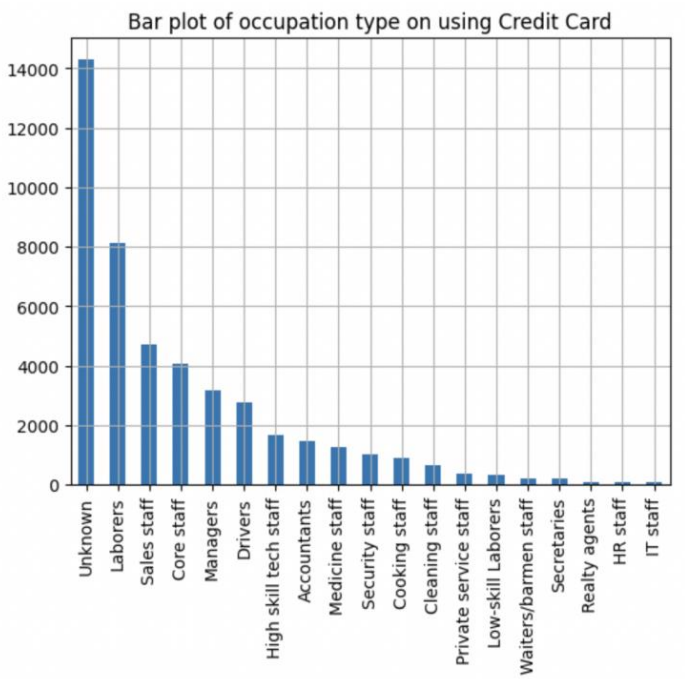
- Data Description:


train_df.describe()

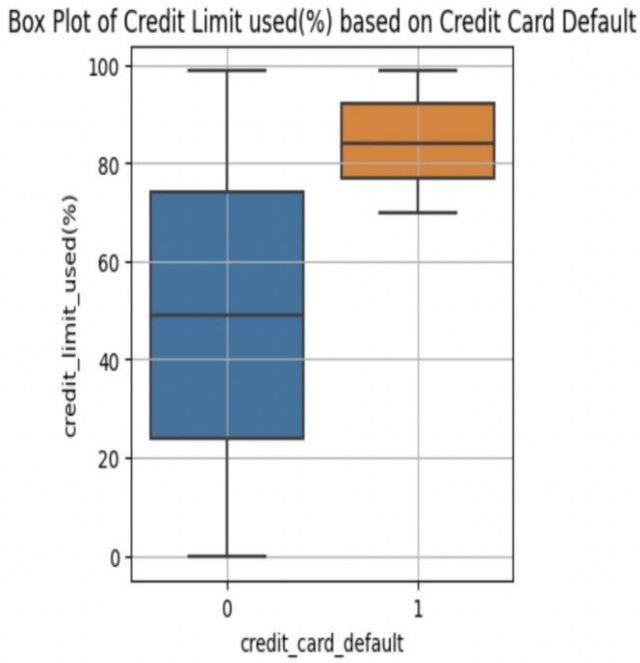
8 rows × 13 columns

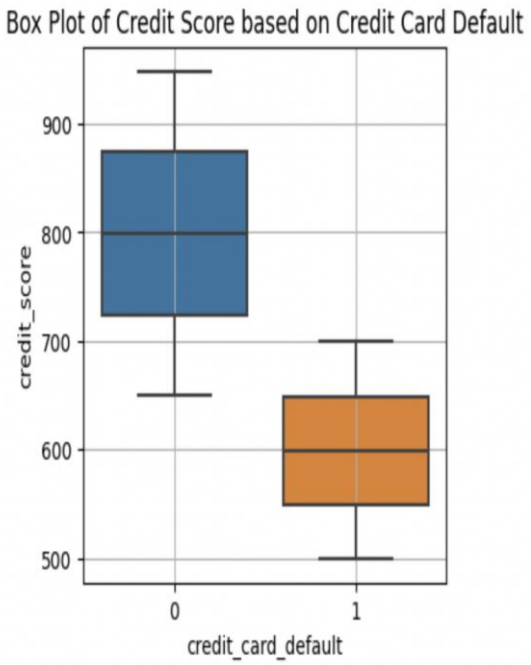
	age	no_of_children	net_yearly_income	no_of_days_employed	total_family_members	migrant_worker	yearly_debt_p
count	45528.000000	44754.000000	4.552800e+04	45065.000000	45445.000000	45441.000000	454
mean	38.993411	0.420655	2.006556e+05	67609.289293	2.158081	0.179111	317
std	9.543990	0.724097	6.690740e+05	139323.524434	0.911572	0.383450	172
min	23.000000	0.000000	2.717061e+04	2.000000	1.000000	0.000000	22
25%	31.000000	0.000000	1.263458e+05	936.000000	2.000000	0.000000	192
50%	39.000000	0.000000	1.717149e+05	2224.000000	2.000000	0.000000	296
75%	47.000000	1.000000	2.406038e+05	5817.000000	3.000000	0.000000	405

- Many High income people are taking higher value Credit loans. We can see failure to correctly predict risk leading high value losses.

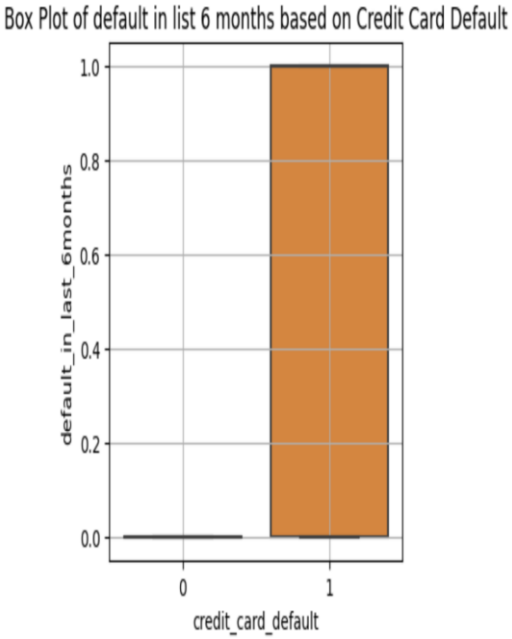


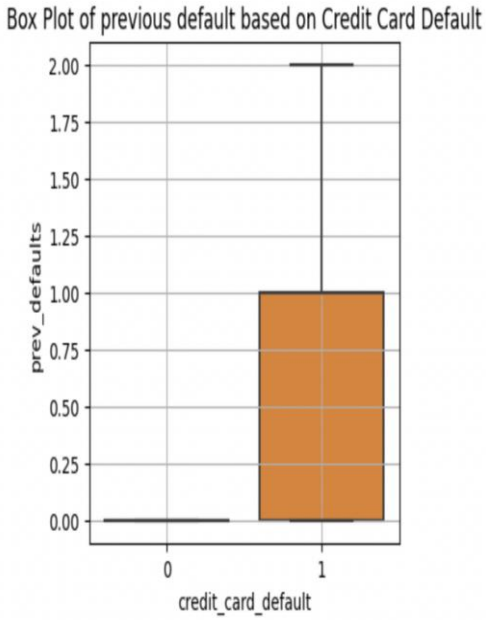
- From below plot, we can see credit defaulter are the one using higher credit value and maintaining low credit score



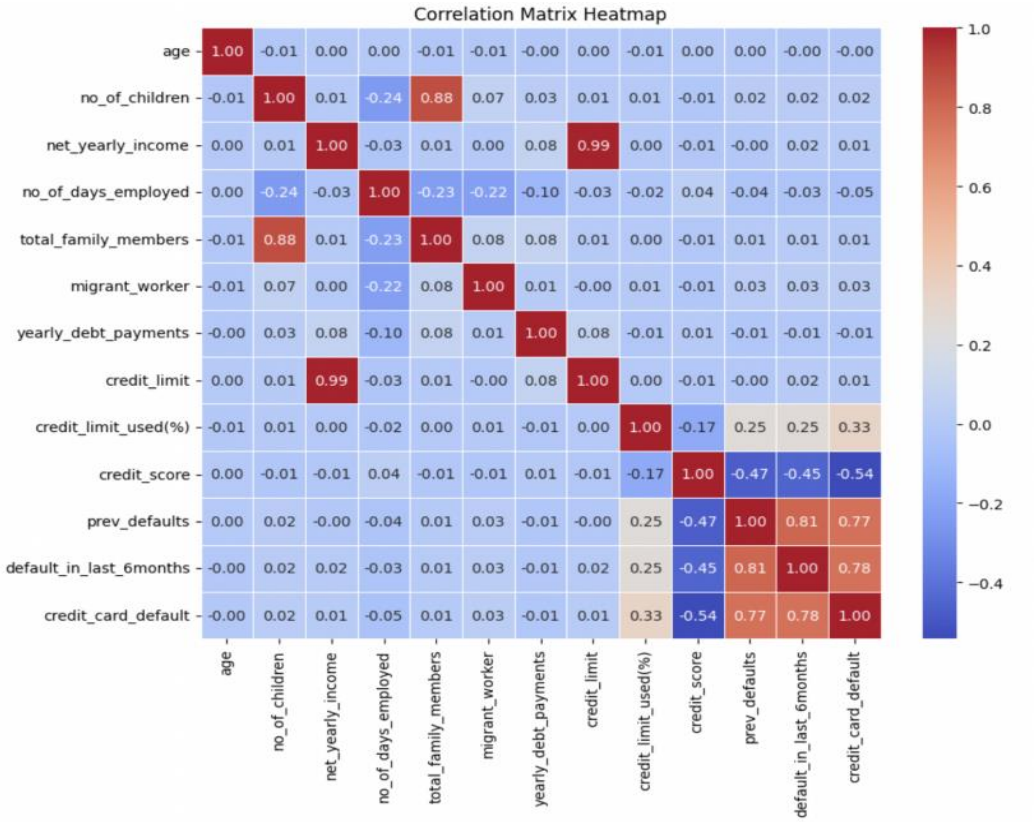


- From the below plot, we can see patterns of previous defaulters having a high chance of defaulting in future providing strong indication of risk.

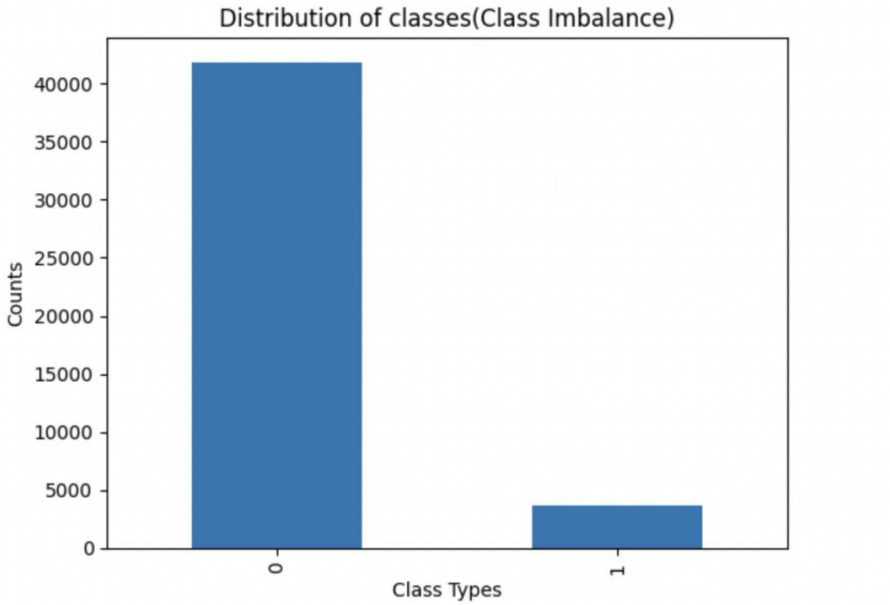




- Correlation plot of values between features and target prediction



- Class Imbalance is present in dataset which is handled using ADASYN



Preprocessing

- **Data Cleaning:** Records with null feature values were dropped and unimportant features such as customer id and date were removed.
- **Outlier removing:** Inter quartile range was used to remove outliers, the data was well within lower bounds but some records with values greater than 1.5 IQR was found which were removed
- **Scaling:** Standard scalar was used as the data contained features of different scales to enforce a mean of 0 and SD of 1.

- **Quantifying Label data:** One hot encoding was used to Quantify categorical variables as they lacked hierarchy.
- **Class Imbalance:** There was a huge class imbalance of 10:1 in non default and default cases. We used ADASYN technique for resampling the training data to improve models ability to predict minority class.
- **Dimension Reduction:** PCA was applied to check the number of principal components, furthermore the data was used with a logistic regression model to compare the accuracy with principal features.

Models and Architecture Details

Advance Models:

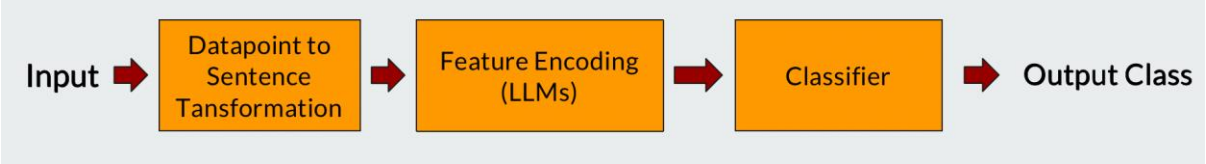
- **Graph Neural Network:** Graph of customer relationship is created using cosine similarity score of threshold greater than 80 percent. Corresponding graph is then passed to Graph Neural Network for more efficient feature representation as well as classifying high risky customers with improved accuracy. Graph Neural Networks improve feature embedding by combining neighbour customers having high similarity to provide more generalised representation, improving gap between decision boundaries of two classes resulting in efficiency classification of clients.

[7]

- **Large Language Models(LLMs):** Natural Language based customer profiling**[1]** is done by transforming data points in the dataset into language representation which can easily be learned by LLMs like BERT and provide mode enhanced representation of feature embedding which can be easily classified using a dense classification layer. We have LLMs trained on a given dataset as the model was able to represent features in terms of more core embedding level taking into consideration relationship between features as well as weighting importance of feature considering large previous history of word relationship as well as improving its classification.
  Feature Transformation(Customer profiling):
  Datapoint is convert into natural sentence for LLMs to provide efficient representation

customer_id	name	age	gender	owns_car	owns_house	no_of_chil	net_yearly	no_of_da	occupation	total_fam	migrant_	yearly_debt_pa	credit_limit	credit_lim	credit_	prev_defaults	default_in_last_	credit_card_default
CST_123268	Sarah Mar	46	F	Y	N	0	252765.91	2898	Accountants	2	1	37046.86	40245.64	19	937	0	0	0

The customer is 46 years old, owning N house and Y car, has 0.0 children, with a net yearly income of $252765.91. They are unemployed for 289 8.0 days, working as a Accountants, and have a total family of 2.0 members. The customer incurs yearly debt payments of $37046.86, has a cred it limit of $40245.64 with 19% utilization, possesses a credit score of 937.0, with some previous defaults, and has defaulted on any credit c ard payments in the last 6 months.



- **Multi Model Ensemble Learning:** In this method, we are combining multiple models such as Logistic Classifier, SVM, Random forest, SGD, leading to learning different data points distribution by different algorithms and effectively providing more generalised results which improves overall performance of credit prediction.[4]
- **Deep Neural Network:** Multilayer deep neural network is used to learn non linear mapping of feature vectors to output. Multiple dense layers leading to learning of more intricate features extraction which passed to final decision boundary leads to improving overall efficiency of credit default classification.

Diversity Models:

- **Naive Bayes(Probabilistics):** It is based on bayes' theorem with assumption of features independence utilising Probabilistics based approach for classification.
- **KNN (Non parametric):** It is a non-parametric model following majority rule by taking neighbourhood points into consideration while classification data points.

Standard Models:

- **Logistics Regression:** Binary classification using logistic function for accurately prediction probabilities of specific class to have efficient decision boundary.
- **SVC**: It maximised the decision boundary between data points of two different classes by transforming data points into representation spaces using various kernels.

Tree Based Models

- **Random Forest:** Combined multiple random decision trees combining different possibilities and providing more generalisation in the final credit decision.
- **XGBOOST**: It is an efficient version of tree model using gradient boosting algorithm for providing overall optimization of trees at each incremental level.

Experiment Results

Below is the table listing various performance metrics such as balanced accuracy, overall f1 considering class imbalance in the dataset.

Algorithms	Train Accuracy	Precision	Recall	Overall/Balanced F1-Sc	Test Accuracy	Methods
Logistic Regression	0.945	0.963	0.946	0.951	0.956	On normalized data with Grid Search and stratified CV
PCA Logistic	0.976	0.976	0.977	0.976	0.893	On normalized data with Grid Search and stratified CV
SVM	0.992	0.969	0.968	0.969	0.922	On normalized and Over sampled data balancing classes
Random Forest	0.99	0.978	0.977	0.976	0.877	On training data with no normalization/ oversampling
Neural Network	0.984	0.96	0.966	0.967	0.934	On normalized and Over sampled data balancing classes
SGD Classifier	0.968	0.963	0.951	0.955	0.945	On normalized and Over sampled data balancing classes
Multimodel Ensemble	0.975	0.964	0.949		0.956	
Naive Bayes	0.968	0.9	0.89	0.96	0.85	On training data with no normalization/ oversampling
XGBoost	0.973	0.89	0.91	0.97	0.89	On training data with no normalization/ oversampling
KNN	0.969	0.9	0.89	0.96	0.83	On training data with no normalization/ oversampling
GNN	0.973	0.9724	0.9729		0.97	On normalized and Over sampled data balancing classes
LLM(BERT)		0.84	0.89	0.97	0.98	

Results Explanation:

- Logistic Regression: Shows strong performance with high training accuracy (0.945), balanced accuracy (0.956), and F1-score (0.951). The model uses normalised data with Grid Search and stratified cross-validation, which is beneficial for generalisation.

- PCA Logistic Regression: training accuracy (0.976) but lower balanced accuracy (0.893) probably due to loss of minor components. The F1-score is excellent (0.976). This model also employs normalised data with Grid Search and stratified CV, emphasising model tuning and validation.
- SVM: Training accuracy (0.98) among all models, with a good F1-score (0.969) and balanced accuracy (0.922). It uses normalized and oversampled data, indicating an emphasis on handling class imbalance and feature scaling.
- Random Forest: Very high training accuracy (0.99) with a high F1-score (0.976), but the balanced accuracy (0.877) is relatively lower. Trained on data without normalisation or oversampling, which might affect performance on diverse datasets.
- Neural Network: High training accuracy (0.984) and F1-score (0.967), with solid balanced accuracy (0.934). The model uses normalised and oversampled data, similar to SVM, suggesting a focus on class balance and feature scaling.
- Weighted SGD Classifier: Good training accuracy (0.968) and F1-score (0.955), with balanced accuracy (0.945). One of the reasons for having high balanced accuracy could be adjustment of weights with gradient descent.
- Multimodel Ensemble: Good training accuracy (0.975) and highest balanced accuracy (0.956). This shows that ensemble provides best balanced accuracy using advantages of multimodal voting .
- Naive Bayes, XGBoost, KNN: These models have moderate balanced accuracy (around 0.85-0.89) and good F1-scores (0.96 and 0.97 for Naive Bayes and XGBoost, respectively). They are trained on data without normalisation or oversampling.
- GNN: Solid training accuracy (0.973) and very high balanced accuracy (0.97). This model also uses normalised and oversampled data, suggesting an effective approach to class balance and scaling.

Supplementary results(Best Parameters):

In most of the models Grid search CV was used to get the best hyper parameter for model.The use of Grid Search CV for hyperparameter tuning across most models shows a defined approach to optimising model performance. Parameter choices, such as regularisation in logistic regression and SVM, tree depth and ensemble size in random forest, and the adjustment of weights in SGD, aligns with the model tuning

The parameter choices were determined using Grid Search CV, optimising for each algorithm's best performance:

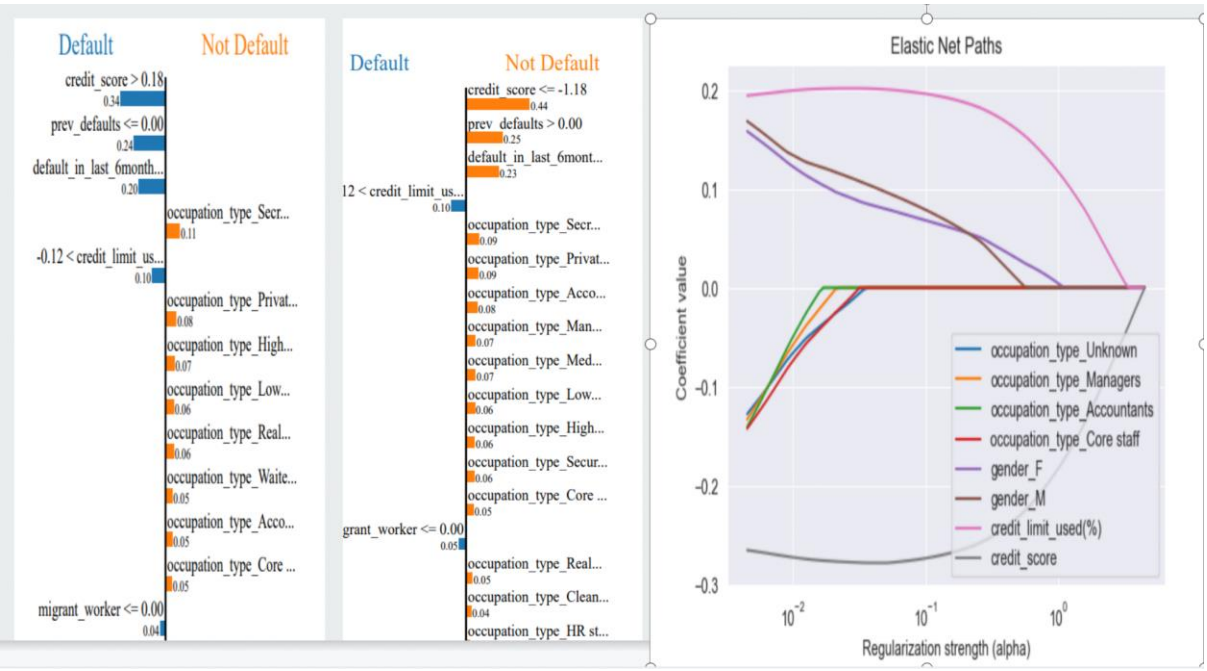
- **PCA with Logistic Regression:**
  Chosen parameters: 'logreg__C': 0.1, 'logreg__class_weight': None, 'logreg__max_iter': 500, 'pca__n_components': 8.
  Balanced regularisation was chosen, maintaining the tradeoff between overfitting and accuracy, the model converged well at 500 iterations retaining 8 principal components for dimensionality reduction.
- **Logistic Regression:**
  Chosen parameters: 'C': 0.001, 'class_weight': 'balanced', 'max_iter': 100.
  Small regularisation was used to avoid overfitting and allow the logistic regression model to generalise well. Further automatic class weight balancing and limited iterations were selected to avoid overfitting.
- **Support Vector Machine (SVM):**
  Chosen parameters: 'svc__C': 10, 'svc__gamma': 'scale', 'svc__kernel': 'rbf'.

A regularization parameter of 'svc__C' = 10 was selected to control the trade-off between smooth decision boundaries and classifying training points correctly.
gamma == 'scale indicates that the scale of the kernel is automatically computed based on the inverse of the number of features.
and kernel = rbf suggests that the model worked best by using non linear kernel, showing non linearity in training data.

- **Random Forest:**
  Chosen parameters: 'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 200.
  Minimum samples per leaf 2 and split 5 were set to avoid overfitting and create a balanced tree. The number of trees 200 in the forest was chosen to build a robust ensemble.

Explainability in credit card default prediction

We used lime to describe the factors involved in predicting class of specific data, and used elastic net paths so visualise coefficient of features at various penalty levels:



COnfirming to above results, Credit score, previous defaults, default in 6 months and credit limit happened to be the most important features in deciding the Default classification for the specific record., whereas occupation had a significant role in classifying the record as not default as shown in the right image. Elastic net path further confirmed credit score, credit limit were more robust to penalty compared to other features.

Discussion

We can draw below inferences from the above discussed models:

- Stratified CV with grid search helped to get the best estimator for models, considering class imbalances. PCA helps to get the minimum number of components (8 in our case) to explain the majority of variance in the data and helps reduce the dimensionality. Though it is not accurate as it discards some of the minor components during dimensionality reduction, emphasising the delicate trade-off between accuracy and complexity. Weighted Logistic regression with best parameters works better in terms of accuracy.
- SVM with non linear kernel (rbf) worked slightly better than linear kernel demonstrating some non linearity in the data. Tree based methods such as Random forest and XG boost offered a balanced accuracy of 87.7% is relatively lower, possibly indicating challenges due to imbalanced dataset because the model is trained without normalisation or oversampling, which might contribute to its performance variation.
- Random forest offers feature importance that confirms presence of major components almost equal to 8 (Confirming PCA results), wherein credit score, credit default, and income are the top features. The weighted SGD classifier exhibits good performance with balanced accuracy of 94.5%. The adjustment of weights during gradient descent likely contributes to the high balanced accuracy, showcasing the effectiveness of this approach.

- **Naive Bayes, KNN:** These models exhibit moderate/poor balanced accuracy (around 85-89%), suggesting inability to perform well on complex data.
- The neural network demonstrates high training accuracy (98.4%), a solid balanced accuracy (93.4%), and an impressive F1-score (96.7%). The multimodel ensemble approach yields good training accuracy (97.5%) and the highest balanced accuracy of 95.6%. Leveraging the advantages of **multimodal voting,** this ensemble method proves effective in achieving a well-balanced model.
- The graph neural network (**GNN**) achieves a solid training accuracy of 97.3% and an exceptionally high balanced accuracy of 97%. Similar to Neural Network, the use of normalised and oversampled data reflects a successful strategy for addressing class balance and feature scaling.
- **LLMs(BERT):** Large Language Models seems to provide more optimised representation which can be easily use for classification of customer profile with accuracy of 98% providing state of the art performance and most efficient any other model in our experiment set.

Conclusion

Credit risk prediction is an important challenge even in today's world because of ever evolving market conditions. Financial institutes are spending **millions of dollars** every year on researchers all over the globe to provide more efficient models. In the project, we have focused on improving credit risk prediction by experimenting with more advanced machine learning models like GNN, LLMs, Multi model ensemble along with standard classification model to provide more efficient prediction of customer default. We found that multi model ensembles improve benchmark of individual standard model accuracy, which is further improved by using Graph Neural Network considering graph networks using higher customer cosine similarity for prediction. At the end, LLMs like BERT seem to optimise feature representation and state of the art performance then any other model.

Future Scope

The strong performance across various models shows the effectiveness of different strategies for addressing class imbalance, feature scaling, and model complexity. Utilising more efficient graphs representation for graph neural networks as well as using different LLMs can provide more efficient in capturing real-world dynamics for credit risk prediction. Alternative techniques based on ensemble methods or attention mechanism based neural network architectures could be considered for further improvement.

References:

1. Yin, Yuwei, et al. “FinPT: Financial Risk Prediction with Profile Tuning on Pretrained Foundation Models.” *arxiv.org*, 2023, https://arxiv.org/pdf/2308.00065.pdf
2. Xiang, Sheng, et al. “Semi-supervised Credit Card Fraud Detection via Attribute-Driven Graph Representation Authors.” *https://ojs.aaai.org/*, 2023, https://ojs.aaai.org/index.php/AAAI/article/view/26702.
3. Huang, Menglan, et al. “Credit Default Prediction Based on Improved Smote Algorithm and GA_LightGBM Algorithm.” *dl.acm.org*, 2022, https://dl.acm.org/doi/10.1145/3570236.3570279
4. Xu, Ronghua, et al. “Credit Default Prediction via Explainable Ensemble Learning.” *dl.acm.org*, 2021, https://dl.acm.org/doi/abs/10.1145/3503181.3503195.

5. Feng, Bojing, et al. “Every Corporation Owns Its Structure: Corporate Credit Ratings via Graph Neural Networks.” *arxiv.org*, 2020, https://arxiv.org/abs/2012.01933.
6. https://www.kaggle.com/code/vishnu0399/amex-credit-card-default-prediction-98-accuracy/input
7. https://link.springer.com/chapter/10.1007/978-3-031-20891-1_44
8. https://www.mathworks.com/help/deeplearning/ug/compare-deep-learning-networks-for-credit-default-prediction.html
9. PPT: