

Student(s)

Elif Ceryansuyu
Tülay Çolakoğlu
Utku Alemdağ

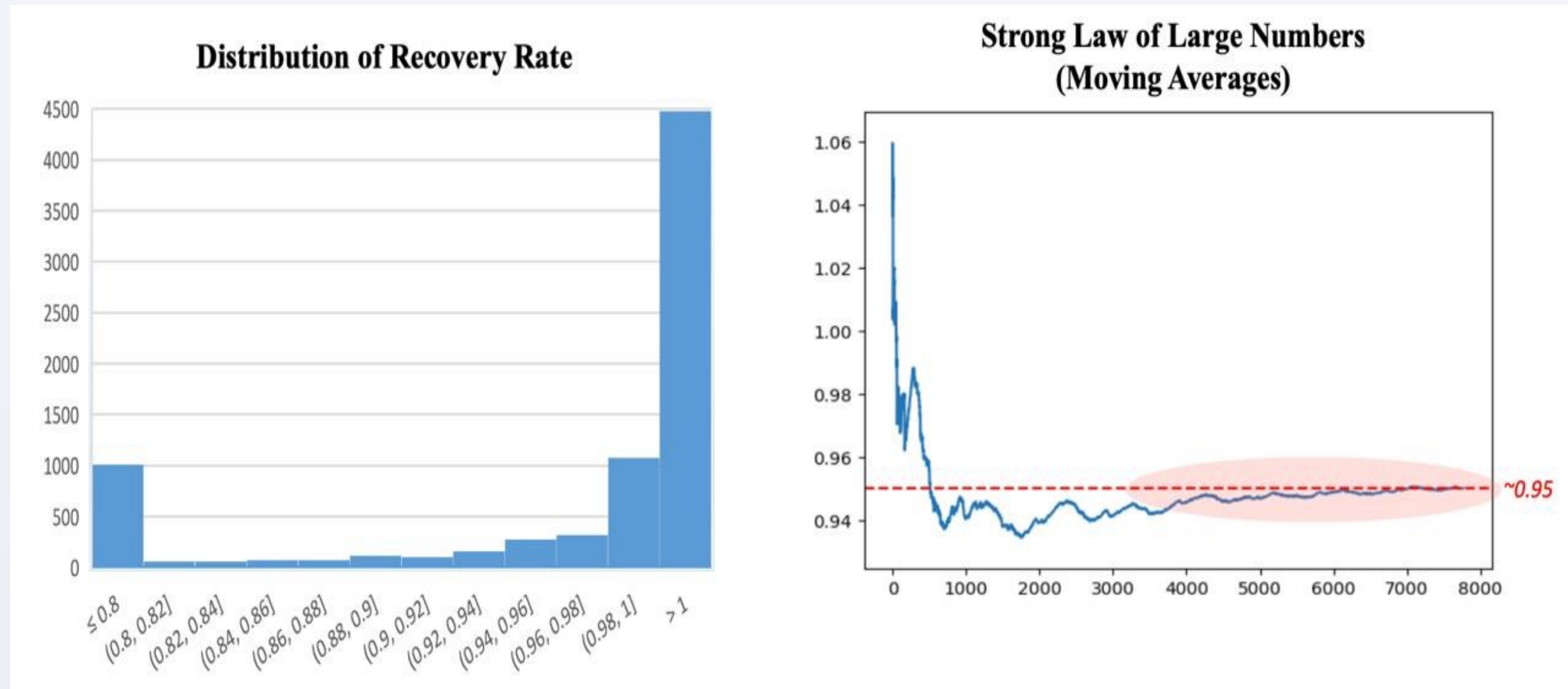
Faculty Member(s)

Hüseyin Özkan

Company Advisor(s)

Şahin Nicat

ABSTRACT

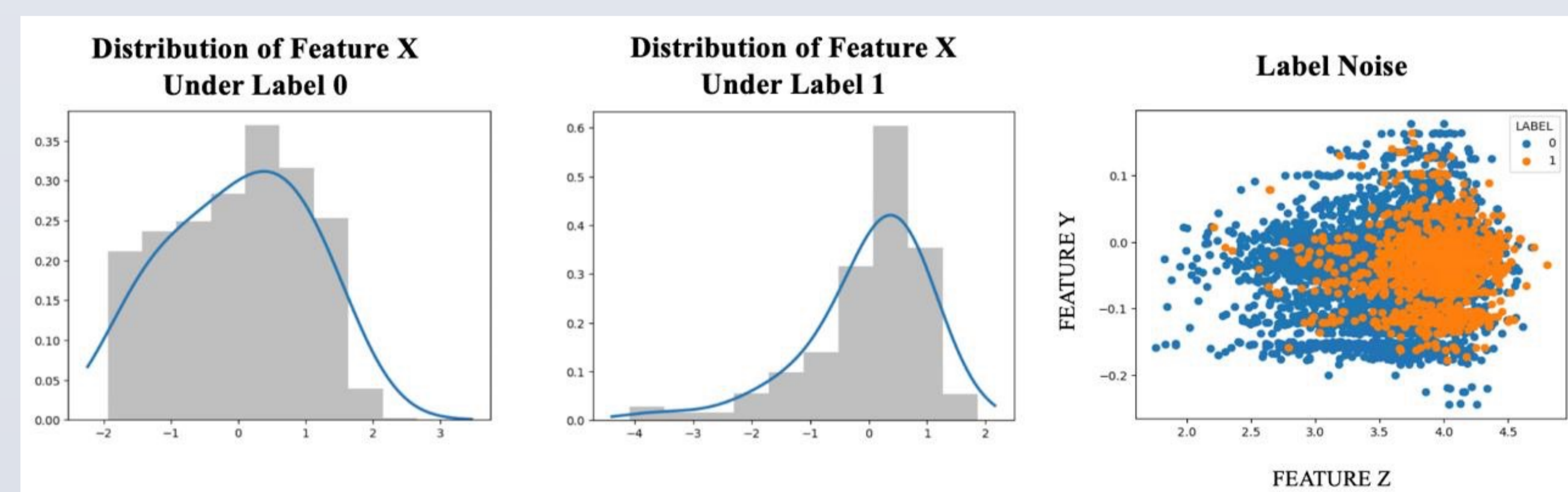


The profitability of financial institutions relies heavily on successful credit risk measurement models. To comply with banking regulations, financial institutions employ three key components for measuring credit risk: the probability of default (PD), loss-given default (LGD), and exposure at default (EAD). However, managing credit risk is a complex task due to its multidimensional nature and intricate interactions. Furthermore, the scarcity of the "default" label creates an imbalanced data problem, which is a common challenge in credit risk measurement models. The project is dedicated to deriving predictive outcomes from imbalanced data through the exploration of algorithm-level and data-level approaches. We made a classification prediction as to whether the customer will pay more or less than 0.80 of the loan he received when a customer who received a loan from Koçfinans is subject to legal proceedings (i.e., presently in default). The model we have established can make predictions at the desired false positive rate, it is possible to take risks and make predictions at the level the company wants. The overarching impact of the present work for a company is identifying customers who will fall short of expectations. In accordance with the company's strategic goals, greater resources may be allocated to such customers, with the aim of improving their performance. Alternatively, resources may be directed towards customers who are expected to perform well, with the expectation that they will yield better returns. By adopting this approach, the organization can optimize its use of limited resources and achieve greater efficiency.

OBJECTIVES

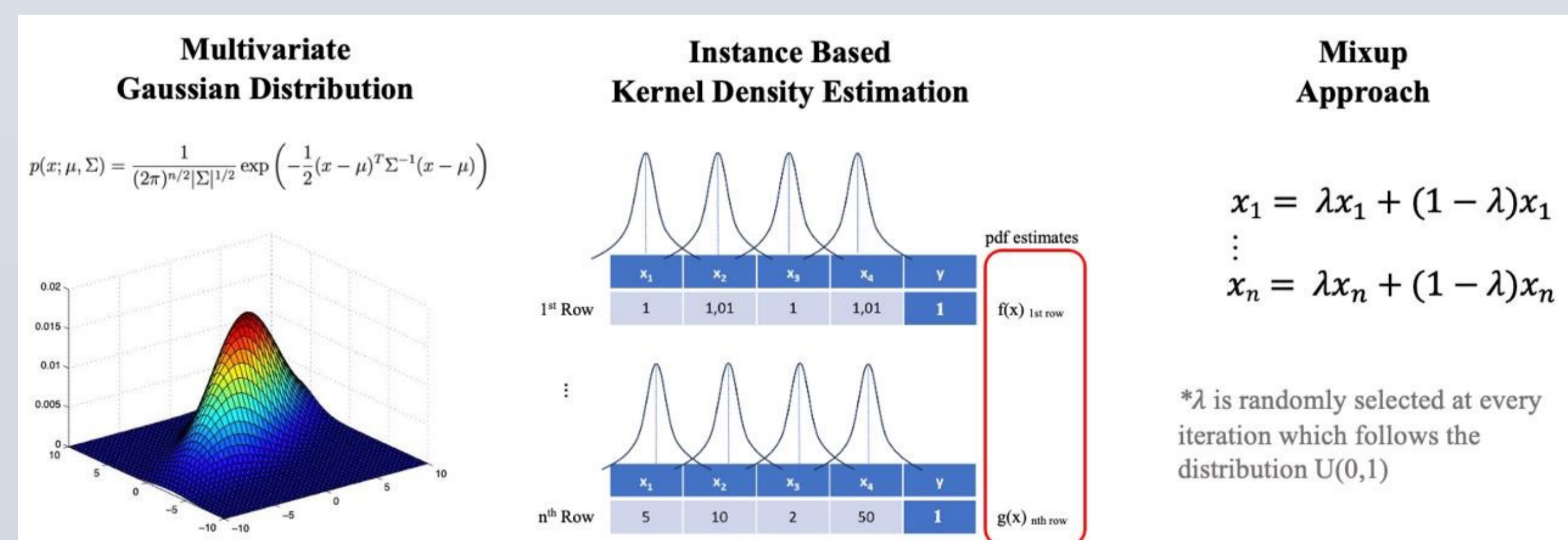
The primary objective of the project is to accurately identify the default contracts that cannot be recovered. To achieve this, we required a dependable algorithm suited to our specific needs. Our goal at this stage is to obtain satisfactory results and then address the class imbalance issue. To do so, we incorporated the Gaussian classifier, Kernel Density Estimator, and Random Forest techniques and incorporated synthetic data generation techniques such as Random Oversampling with Multivariate Normal Distribution, Random Oversampling with instance-based KDE, and Mixup approach. We took great care in selecting the features, adjusting the algorithm's parameters, and choosing the appropriate normalization technique for each method.

PROJECT DETAILS



The target variable of the dataset provided by Koçfinans is 'recovery rate' which determines the proportion of loss that is retrieved from customers in default. The dataset contains 7830 customers and the recovery rate is not bounded within the interval [0,1].

We have incorporated feature selection techniques to identify and retain the most relevant features for modeling purposes. This process involves analyzing each feature's contribution to the model's performance and selecting only the most significant features to reduce the dimensionality of the data. By doing this, we can improve the efficiency and interpretability of our models and ensure that they perform well on new, unseen data. After examining the features, we decide to move on with 13 of them, the most significant ones among others. Our findings indicate that the significance of features is detectable through the disparity between the distribution of feature X under label 0 and that under label 1, thereby serving as a prominent leading indicator for feature selection. Therefore, greater disparity in distributions results in a more substantial contribution to the models' performance.



To address the issue of data imbalance, we utilized oversampling methods. This involved augmenting the number of minority class samples by replicating them using various techniques. One such method involved oversampling from a multivariate normal distribution, utilizing the mean and standard deviation of each feature to randomly generate new data from the same distribution. Additionally, we employed kernel functions with Gaussian kernels to fit each row of the training set for the minority class, allowing us to sample new data that adhered to the distribution of each specific row. Furthermore, we implemented the Mixup technique, a data augmentation method involving the linear interpolation of pairs of rows to create new samples. During the selection of pairs, a subset consisting of points on the edges was utilized. The training set was employed to test the model to identify points that were challenging to classify. To accurately establish this subset of points, the k-nearest-neighbor method was utilized to eliminate potential outliers that might confuse the model.

CONCLUSIONS

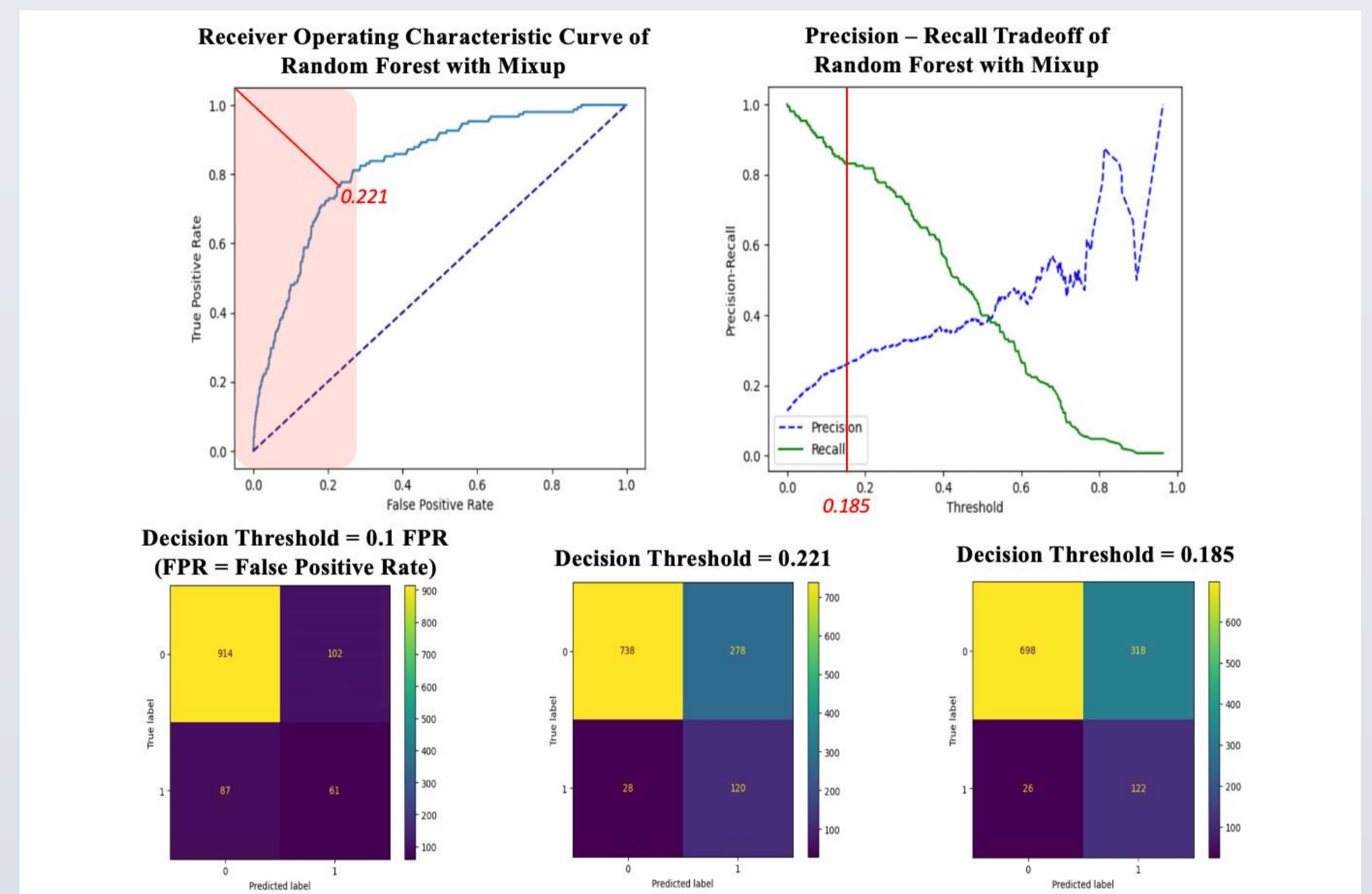
			AUC Scores of Models*			
			Without Synthetic Data Generation	Random Oversampling with Multivariate Normal Distribution	Random Oversampling with instance-based KDE	Mixup
Parametric Approach	Multivariate Normal Distribution	Bayes' Classification	0.7717	-	0.7338 <i>0.8934</i>	0.7813 <i>0.8553</i>
Non-parametric Approach	Kernel Density Estimation where Gaussian Kernel Function is used	Bayes' Classification	0.7988	0.818 <i>0.95</i>	-	0.8173 <i>0.951</i>
Tree Based Algorithm	Random Forest	Classification	0.8198	0.7748 <i>0.977</i>	0.8136 <i>0.99</i>	0.826 <i>0.99</i>

* Red scores denote training scores, while black scores denote test scores.

The aim here is to choose the best model and data generation technique duo (that is, the model and data generation technique which has the highest AUC) to improve learning as much as possible from the imbalanced dataset and make the most accurate prediction. The AUC scores for all three models with and without synthetic data generation are displayed in the table above.

During the process, we have encountered limitations, such as issues related to data quality. The differential classification of two similar customers due to label noise can diminish the predictive power of machine learning models, effectiveness of synthetic data generation and exacerbate data imbalance issues. Consequently, it is recommended that the initial step in the project's continuation involves the identification of similar customers in the dataset and the reclassification of customers labeled as 0 to 1. This approach aims to solve the data imbalance problem and facilitate the identification of customers already classified as 1.

It appears that overfitting was encountered in the Random Forest model, as indicated by the AUC scores in the training set. The best performing model seems to be a combination of Random Forest and Mixup, although the Mixup and KDE duo also demonstrates similar AUC scores with a lower overfit rate, making it a preferable choice.



The ROC curve of the Random Forest with Mixup data generation method model on the upper left graph depicts effective performance at True Positive Rates without a significant increase in the False Positive Rate at low thresholds.

The model's performance at different thresholds was assessed using the AUC score, and optimal thresholds were determined using various methods. One method involved selecting the threshold closest to the top-left corner of the ROC curve, while the other method considered the tradeoff between precision and recall to select the highest possible values.

Confusion matrices for the calculated thresholds are presented, with the observation that different models can be utilized by selecting different thresholds based on the prioritized metrics.

To advance the project, our recommendation is to direct attention towards assessing the predictive power of machine learning models, evaluating the efficiency of synthetic data generation, and addressing issues related to data imbalance with handling label noise and enhance discrimination power of feature space.

REFERENCES

Albanesi, S., & Vamosy, D. (2021). Predicting Consumer Default.

Alonso-Robisco, A. & Carbo, J. M. (2022). Can machine learning models save capital for banks? Evidence from a Spanish credit portfolio. *International Review of Financial Analysis*, 84.

Brown, D. T., Ciochetti, B. A., & Riddiough, T. J. (2006). Theory and evidence on the resolution of financial distress. *The Review of Financial Studies*, 19(4), 1357–1397.

Hartmann-Wendels, T., Miller, P., Töws, E. (2014). Loss given default for leasing: Parametric and nonparametric estimations. *Journal of Banking & Finance*, 40, 364–375.

Koç, U., & Sevgili, T. (2020). Consumer loans' first payment default detection: a predictive model. *Turkish Journal of Electrical Engineering and Computer Sciences*, 28(1), 167-181.

Li, P., Qi, M., Zhang, X. (2016). Further investigation of parametric loss given default modeling. *Journal of Credit Risk* 12(4): 17-47.

Li, P., Zhang, X., Zhao, X. (2018). Modeling Loss Given Default. *Federal Deposit Insurance Corporation and Centre for Financial Research Working Paper Series*.

Mancisidor, R. A. (2020). *Deep generative models in credit scoring*. [PhD dissertation, The Arctic University of Norway].

Qi, M., X. Zhao. (2011). A comparison of methods to model loss given default. *Journal of Banking and Finance* 35, 2842-2855.

Sigrist, F., & Hirschsall, C. (2019). Grabit: Gradient tree-boosted Tobit models for default prediction. *Journal of Banking and Finance*, 102, 177-192.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., Asadpour, M. (2020). Boosting methods for

Tuovila, A. (2023). Risk-Weighted Assets: Definition and Place in Basel III. Retrieved from <https://www.investopedia.com/terms/r/riskweightedassets.asp>.