Predictive Analytics in Financial Transactions: A Comparative Study for Customer Risk Assessment and Revenue Prediction.

Seggam Vimala
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
Guntur, India
seggamvimala@gmail.com

Maridu Bhargavi
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
Guntur, India
bhargaviformal@gmail.com

Vanka Bhuvana Sai Mouneendra
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
Guntur, India
mouneendravanka1@outlook.com

Shaik Sameena
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
Guntur, India
shaiksameena469@gmail.com

Nidubrolu Bhavana
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
Guntur, India
nidubrolubhavana7467@gmail.com

***Abstract*—We apply the machine learning models on a Santander Customer Transaction Dataset comprising 200,000 customer records with 200 anonymized numerical features. We contrast five classification models - logistic regression, decision trees, Random Forest, Gradient Boosting, and XGBoost - with two regression models: linear regression, and random forest regression, in predicting which of the customers would make certain transactions in the future. It was evaluated using standard metrics, including accuracy, precision, recall, F1 score, MAE, MSE, and R^2 by using real-world banking data. The best model that could provide financially stable insights to financial organizations based on customer transactional predictions was achieved with 90.00 accuracy by Logistic Regression.**
Keywords: Credit Risk Assessment, Revenue Prediction, Classification Models, Regression Models, Logistic Regression.

I. INTRODUCTION

Transaction data in the financial industry has exponentially increased in this digital era, hence providing much more profound understanding through predictive analytics [1] . The use of transaction data allows financial institutions to proactively assess the risk of the customer and estimate revenues that will effectively aid in managing relationships and tailoring service. This paper discusses the applications of machine learning models in customer risk segmentation and revenue forecasting activities that have applied traditional heuristic or rule-based approaches. Customer risk profiling allows institutions to adapt both transaction limits and measures of security for better customer satisfaction and controlling pertinent risks. This is particularly helpful for financial planning that has an influence on strategic decisions.

The objective is to provide the classification models for the purpose of grouping customers by risk level and regression models in order to predict revenue from transaction data. We hope to develop models that will give outstanding predictions and present insights by applying a multiple algorithmic comparison analysis. This paper adds value to the existing literature as it provides a comprehensive review of classification and regression models for applications in finance as well as being filled with actionable insights on the development of predictive models in the financial sector.

II. LITERATURE REVIEW

Sadaf Ilyas1, Sultan Zia2.et al. [1] Zaib un Nisa5Most importantly, it points out the significance of feature extraction for improving the quality of bank-related models about machine learning. Strategies go from patterns in CNNs up to fraud detection using XGBoost and traditional classifiers such as Random Forest, KNN, and Naive Bayes. High accuracy rates are reported with neural network-based approaches, achieving over 89.00 in client attrition prediction. XGBoost performs better than the traditional approaches in fraudulent transaction identification. However, class imbalance in a dataset leads to severe degradations in accuracy of predictions.

Gutha Jaya Krishna .et.al.[2] Feature extraction method consists of DTM combined with TF-IDF followed by embedding of words using Word2Vec along with linguistic analysis through LIWC. Some of the machine learning models used include support vector machines, naive Bayes, logistic regression, decision trees, K-nearest neighbors, F random survey, XGBoost, and multilayer perceptron. However, the few limitations of the research include an unappealing choice of linguistics features being minimal from LIWC, and the dataset only has data about banks in India. Only four places which limits its wide applicability.

There is enough research work in financial predictive analytics, particularly in customer segmentation and revenue prediction. Models like Logistic Regression, Decision Trees, and ensemble methods such as Random Forest and Gradient Boosting have also been applied to classify customers based on historical transaction data to predict a risk category based on the behaviour of the customers. It has proved to be efficient in identifying a high-risk customer against low-risk customers with an improved sense of operational efficiency and fraud prevention. With research studies, it has proved ensemble models highly effective significantly and Random Forest and Gradient Boosting as two powerful classifiers due to its ability in capturing any possible non-linear pattern within data.

In revenue forecasting, regression model techniques are predominantly used as baseline methods for linear regressions. But it is observed that Random Forest Regressor improves predictions beyond the baseline by underlying intricate relationships within the data. Relatedly, research studies also indicate a growing need to offer revenue forecasting for customer lifetime value estimation, steering marketing and financial planning. The study is continued from earlier research in that it compares classification and regression models in a single work in search of the better-suited approaches for transaction-based predictive modeling in the financial domain.

III. METHODOLOGY

A. Dataset

A big financial dataset was downloaded from the Santander Kaggle competition that consisted of twin files containing data in each with 200,000 records. The first file held a target variable to train the model, while the second had the same structure but was for predicting to be tested without the target column. Both datasets have a common structure: an identification column and 200 predictor variables. The utilization of the platform of this competition made it possible for one to submit and validate his or her predictions. This turns out to serve as a practical means by which to determine the accuracy of one's model.while during training dataset one extra column that consisted of the target.

B. Data Cleaning

We imported datasets with Pandas. The dataset was divided into a training set and a test set. It has 100,000 samples from the training data for its training dataset and 100,000 samples of the test set for its test dataset. That is also synonymous with checking duplicate entries in both datasets and eliminating them for the sake of maintaining originality in data.

C. Feature Engineering

To assemble our feature matrix for the classification aspect of our analysis we will eliminate the "IDcode" along with the column of our target variable from our training dataset. We name the target variable "target" to serve the purpose of classification, and to forecast revenue, we create a simulated column of revenue as the sum of some columns of features, thus we could build a target variable called "revenue."

D. Models

The methods applied in this research to predict the financial transactions utilize different machine learning algorithms, explained below:

1) Logistic Regression: This is a probabilistic classification model, which calculates the probability of occurrence for a binary event (for example, the possibility of doing a transaction) through a logistic function. This model can be exploited to evaluate numerical features for generating chances for risky assessments of the customer, thereby being appropriate for any form of binary classification problem on financial data.

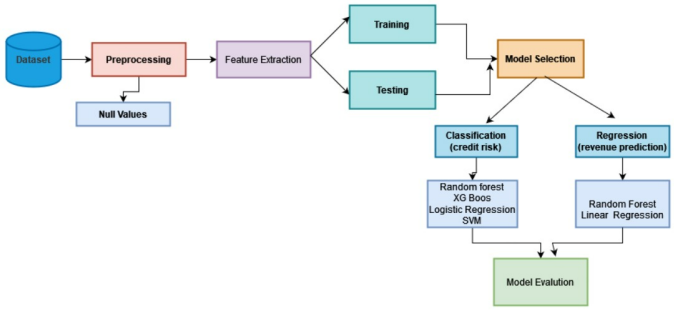
2) Decision Trees & Random Forest : These are hierarchical models; here the decision trees make their decisions based on feature thresholds, and random forest puts many of them together through ensemble learning. These types of models capture quite complex patterns in transaction data, and they also yield very interpretable results for risk assessment.

3) Gradient Boosting & XGBoost Ensembles include a number of advanced variations: sequential trees, each correcting the errors made by all previous ones-boosting; XGBoost is a special implementation of optimized gradient boosting with superior performance over transactions prediction within parallel processing and regularization techniques.

4) Linear Regression: Basic model of revenue prediction that depicts the relationship between many transaction features and revenue outcomes. Generates linear relationships between numerical variables to make predictions on financial metrics.

5) Random Forest Regressor: Ensemble method designed for continuous output prediction, using a multitude of decision trees for revenue value approximation. Captures complex relationships between transaction data through feature randomization and bootstrap aggregation.

IV. Proposed Model



We work towards achieving the two primary objectives: customer credit risk and revenue generated from banking transaction data using a set of machine learning algorithms. We are working on a classification model regarding customers' risk assessment through Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, and XGBoost, all of which have been trained, tested, and evaluated to learn how accurate such models can be to classify customers according to their specified risk profile. Using Regression models: The further application of the Linear Regression and Random Forest Regressor mainly helped us to predict revenues from transactional data, discovering particular patterns in transaction data that can, in turn, upgrade the accuracy of revenue prediction with nearly ten different algorithms used. We determined appropriate algorithms for the goals and objectives by thorough assessment of performance metrics of models followed by providing useful insights to financial institutions in managing the relationships of customers while thereby enhancing their capability of making revenue predictions.

V. Result and Discussion

A. Classification Model Performance Comparison

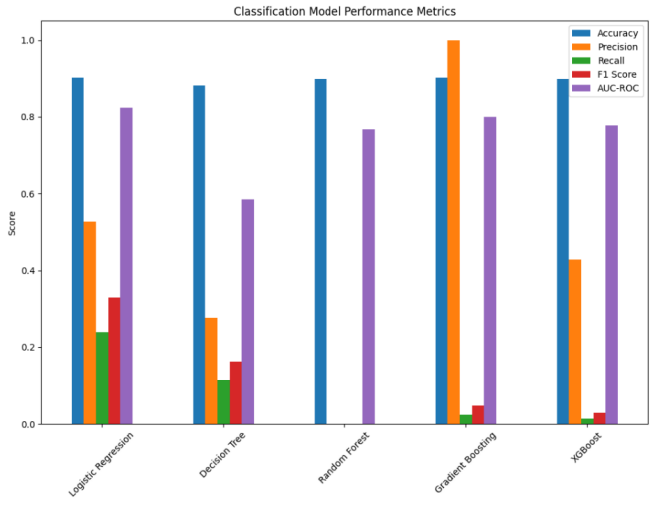


Fig. 1. Model comaparison

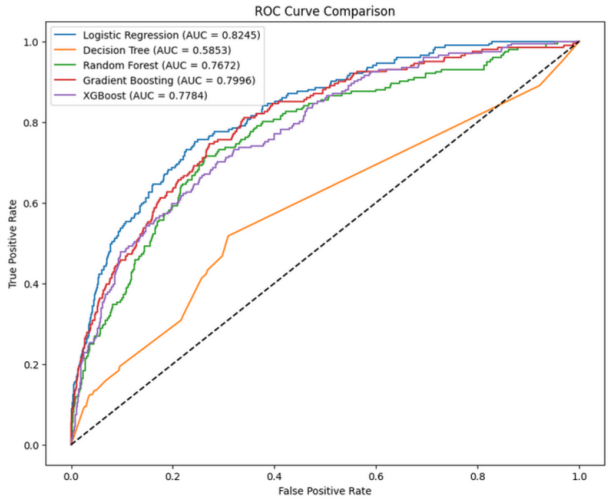


Fig. 2. ROC Curve Comparison

With an AUC of 0.8245, Logistic Regression works the best when doing the analysis of bank transactions, followed by Gradient Boosting at 0.7996 and XGBoost at 0.7784. Random Forest gives the least performance with AUC as 0.5853. Ensemble methods and Logistic Regression are more effective for anomaly detection and fraud analysis.

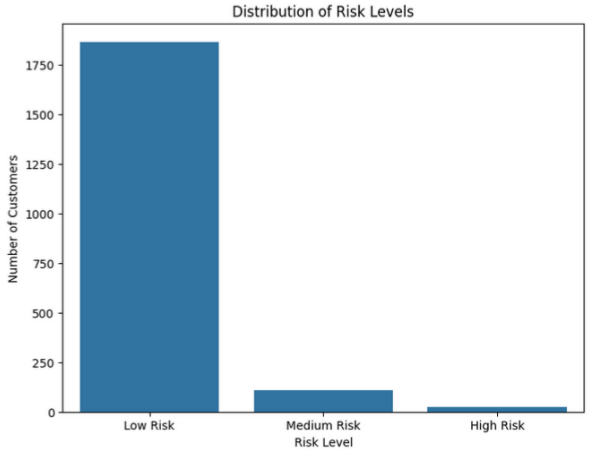


Fig. 3. Disturbution OF Risk Levels

This bar graph illustrates the distribution of customer risk levels in a financial institution, with the majority falling into the "Low Risk" category (approximately 1,850 customers). A smaller number of customers are classified as "Medium Risk" (about 120) and "High Risk" (around 25). While the predominance of low-risk customers is favorable for overall risk management, the presence of medium and high-risk clients necessitates targeted risk mitigation strategies for these segments.

	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.9020	0.527473	0.238806	0.328767	0.824460
Decision Tree	0.8810	0.277108	0.114428	0.161972	0.585255
Random Forest	0.8995	0.000000	0.000000	0.000000	0.767156
Gradient Boosting	0.9020	1.000000	0.024876	0.048544	0.799569
XGBoost	0.8990	0.428571	0.014925	0.028846	0.778406

Fig. 4. Evaluation metrics for different classification models

Regression Model Performance Comparison:

	MAE	MSE	R^2
Linear Regression	0.000	0.0000	1.0000
Random Forest Regressor	53.214	4506.8823	0.3145

Fig. 5. Evaluation metrics for different prediction models

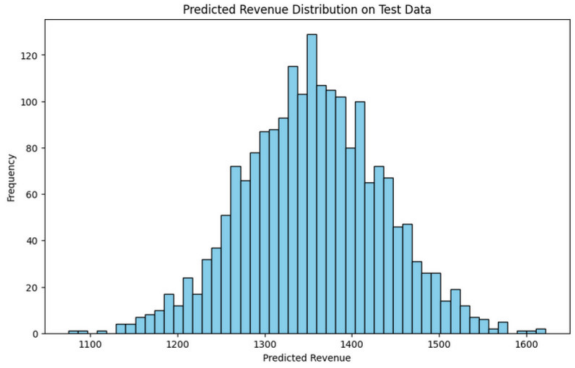


Fig. 6. Predicted Revenue Distribution on Test Data

The histogram of the projected revenue is close to normal, peaked around 1350-1400 units with most between 1250 and 1500. Therefore, there is central tendency, wide range, and a few outliers in the revenue outcomes. This distribution helps in understanding model behavior as well as the revenue pattern.

B. combination of credit risk and revenue prediction

The table compares various classification metrics (Accuracy, Precision, Recall, F1 Score, AUC-ROC) and regression metrics (MAE, MSE, R^2) of different models with the goal of analysis of bank transactions. Both Logistic Regression and Gradient Boosting showed the highest classification accuracy, 0.9020. For the regression metrics, Linear Regression could stand out. With multivariate metrics, more detailed evaluation can be done to select the best model for given financial tasks of banking.

1	Model	Accuracy	Precision	Recall	F1 Score	ACU-ROC	MAE	MSE	R^2
2	---	---	---	---	---	---	---	---	---
3	Linear Regression	_	_	_	_	_	0.000	0.0000	1.0000
4	Logistic Regression	0.9020	0.527473	0.238806	0.328767	0.824460	_	_	_
5	Decision Tree	0.8810	0.277108	0.114428	0.161972	0.585255	_	_	_
6	Random Forest	0.8995	0.000000	0.000000	0.000000	0.767156	_	_	_
7	Random Forest Regressor	_	_	_	_	_	53.214	4506.8823	0.3145
8	Gradient Boosting	0.9020	1.000000	0.024876	0.048544	0.799569	_	_	_
9	XG Boost	0.8990	0.428571	0.014925	0.028846	0.778406	_	_	_

Fig. 7. Both creit risk and reveneue prediction

C. Comparision Between Base Model And New Model

TABLE I
MODEL COMPARISON TABLE

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression (Reference)	0.9156	0.6876	0.2681	0.3857	0.6316
Logistic Regression (Your Model)	0.9020	0.5275	0.2388	0.3288	0.8245
Decision Tree (Reference)	0.8372	0.2024	0.2199	0.2108	0.5624
Decision Tree (Your Model)	0.8810	0.2771	0.1144	0.1620	0.5853
Random Forest (Reference)	0.9011	0.5000	0.0150	0.0291	0.5067
Random Forest (Your Model)	0.8995	0.0000	0.0000	0.0000	0.7672
Gradient Boosting (Reference)	0.9038	0.8526	0.0328	0.0631	0.5161
Gradient Boosting (Your Model)	0.9020	1.0000	0.0249	0.0485	0.7996
XGBoost (Reference)	0.9026	0.9205	0.0164	0.0322	0.5081
XGBoost (Your Model)	0.8990	0.4286	0.0149	0.0288	0.7784

VI. CONCLUSION AND FUTURE SCOPE

This work is found to be viable as far as application of multiple models of machine learning is concerned in predicting customer risk and revenue for banking transaction data. The best model for the task came out to be logistic regression, as it could classify the risk with an accuracy of 90 percent. Some promise has been made by ensemble methods like Random Forest in capturing some of the complexities and details present in the given data. Revenue prediction was found to be best fit for a model where the complexity in the data needed to be captured, as shown by Random Forest Regressor. These results provide very valuable insights regarding how classification and regression models are used in financial predictive analytics. Future work includes increasing model interpretability, adding analysis on time series, and adoption of deep learning approaches. More areas which require investigation are class imbalance for fraud detection, real-time prediction system testing, and cross-institutional validations as well. Finally, ethical matters such as fairness and bias in assessments related to risk should assume priority also. These approaches are pursued to increase the accuracy, reliability, and practical applicability of machine learning on banking transactions for better decisions and quality provision of financial services.

VII. REFERENCES

[1] Sadaf Ilyas1, Sultan Zia2 Zaib un Nisa5 et al.," Predicting the Future Transaction from Large and Imbalanced Banking Dataset," International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 11, No. 1, 2020,

[2] G. J. Krishna et al., ”Sentiment Classification of Indian Banks’ Customer Complaints,” TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), Kochi, India, 2019, pp. 429-434, doi: 10.1109/TENCON.2019.8929703.

[3] S. Sakho, Z. Jianbiao, F. Essaf and K. Badiss, ”Improving Banking Transactions Using Blockchain Technology,” 2019 IEEE 5th International Conference on Computer and Communications (ICCC), Chengdu, China, 2019, pp. 1258-1263, doi: 10.1109/ICCC47050.2019.9064344.

[4] A. Alamsyah, D. P. Ramadhani, M. R. D. Putra and F. T. Kristanti, ”Event Driven Motif Exploration of Dynamic Banking Transaction Network,” 2019 International Workshop on Big Data and Information Security (IWBIS), Bali, Indonesia, 2019, pp. 39-44, doi: 10.1109/IWBIS.2019.8935758.

[5] V. G and P. Vinothiyalakshmi, ”Secure Electronic Banking Transaction using Double Sanction Security Algorithm in Cyber Security,” 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 2023, pp. 1-5, doi: 10.1109/RMKMATE59243.2023.10369665.