Unlocking Business Potential: A Comprehensive Study on Credit Card Defaulter Prediction, Bike Sharing Demand prediction, and Market Basket Analysis with Machine Learning Models

Komal Prakashchandra Patra
M.Sc. Data Analytics
National college of Ireland
Dublin, Ireland
x22210369@student.ncirl.ie

***Abstract*—This research is based on a Machine Learning model which used various different models on 3 datasets. SMOTE technique is performed in order to handle the imbalanced data. The first dataset is about credit card defaulters' prediction, in which it predicted which consumers come under defaulters using Logistic Regression, SVM, Random Forest, and Decision tree. Random Forest outperformed with a better overall accuracy of 90% as compared to other models. The second dataset consists of Bike sharing demand prediction in which the prediction has been made for the count of rental bikes. KNN and Linear Regression were used, where KNN achieved a higher accuracy of 87%. The third dataset was transaction data for Market basket analysis to create a recommendation engine for the products to increase their overall sales.**

***Index Terms*—Market Basket Analysis, Accuracy, precision, Recall, Linear Regression, Random Forest, Apriori Algorithm, Support Vector Machine**

I. Introduction

Credit card is the most critical factor in financial sector that employs ML techniques and data analytics to identify the credit risk. By analysing historical transaction data, the predictive models can be developed to improve the risk and contributes to the stability of the credit card sector. The bike demands go on increasing depending on various attributes in an aim to forecast the future demand accurately. Market basket Analysis is powerful technique in retails sector to recommend the products in order to increase their demands and product sales. Predictive analytics in the retail industry also helps to manage the product promotions and inventory management. The objective is to enhance the shopping experience of consumers and drive the overall business profitability.

"Can ML outperform traditional statistical methods for forecasting the credit card defaulters, especially when using Account and the past bill history as the predictive features?"

The Financial organisation totally depend on predicting the credit card defaulters in order to mitigate the risk. Traditionally it has been calculated mathematically, however ML techniques have made this process automated. The use of account information, encompassing variables such as credit utilisation, their balance limit, pattern of the transactional payments contributes the approach to captain individual's financial pattern. The focus is to evaluate the effectiveness of leveraging the information to uncover the insights. Several machine learning models can be applied in order to predict credit card risk with the model having overall better accuracy.

"What are the features that are affecting the count of rental bikes, and how can these factors be leverage in order to predict the rental bike count accurately?"

There are various factors that are influencing the count of rental bikes. So, in order to understand this element is critical for accurate predictions. The key elements contributing to bike rental count include wind speed, time of the day, day whether its weekend or weekday and various special events. Depending on the time and day of the week, holidays and working days may lead to deviations for the rental bike. ML models such as regression model and time series analyses can be able to obtain the relationships between identified factors and rental bike count.

"What are the patterns and the associations that exist in the consumer pattern behaviour and how can these Market Basket Analysis be employed to recommend the items effectively, optimizing and recommending based on the frequently items purchased?"

MBA helps to uncover the patterns of consumer behaviour, in order to improve the effectiveness of the recommendations. Finding out the relationship between the items that are frequently purchased by the customers. The primary goal is to recognise the patterns in consumer transactions that go beyond the personal product preferences by learning the co-occurence or count of the items. Using MBA techniques, a strong recommendation system that can make suggestions for the items was developed.

II. Related Work

Research in the field of credit card risk prevention via predicting the defaulters has taken various forms over time.

Ruirui Zhang and Ying Chen [1] carried out a study in this area using the BP Neural Network and the k-means SMOTE algorithm. The SMOTE technique is used to handle the imbalance data which occurs when one class in a dataset has substantially fewer instances than another class, which is known as class imbalance. Additionally, they have employed five distinct models, which include Random Forest, SVM, KNN, and Decision Trees. Outperforming all other models, the BP neural network performs better. The proposed algorithm significantly improves the model's prediction performance, as demonstrated by the experimental results, which yield an AUC value of 0.929.

TABLE I
AUTHORS AND METHODOLOGIES USED

Authors	Methodologies Used
Ruirui Zhang and Ying Chen	SMOTE technique, Random Forest, SVM, KNN, and Decision Trees
Liu R	Feedforward neural networks, LSTM, decision trees, KNNs, and random forests
Yuhan Ma	XGBoost
Bacova A and Babic F	Random Forest, XGBoost, Gradient Boosting, and Adaboost
Arora, Bindra, and Kumar Nassa V	KNN, Decision Tree, Random Forest Classifier, Logistic Regression, SVM, PCA, and Naïve Bayes

Another study from Liu [2] focuses on using multiple models, including feedforward neural networks, LSTM, decision trees, K-nearest neighbors (KNNs), and random forests, to predict credit card default. Additionally, they have experimented with dropout, a regularization strategy used in neural networks to avoid overfitting and increase accuracy. The accuracy achieved by the other traditional machine learning models was limited to 40-50%. With a 0.1 dropout rate and a sigmoid activation function, the neural network and LSTM have an approximate 82% accuracy rate.

XGBoost model is used in Yuhan Ma [3] research study, which puts forth the theory that the default normal ratio has an expected score that is doubled. The most significant factors impacting the likelihood of a default credit score are also provided by the AUC of 0.779, which allows for differentiation.

Through the use of Random Forest, XGBoost, Gradient Boosting, and Adaboost, Bacova A and Babic F [4] tried to provide the important ML algorithms to predict the credit card defaulters. All of the features from `PAY AMT 1` through 6 were added to the column `Pay`, and the same was done for `BILL AMT`. The Gradient Boosting algorithm yielded the best results in terms of AUC, but the Bagging algorithm achieved 72% accuracy when taking into account the best precision.

Another takeaway from Arora, Bindra, and Kumar Nassa V [5] is that they employed five distinct models: KNN, Decision Tree, Random Forest Classifier, Logistic Regression, SVM, and Naïve Bayes. Additionally, they have checked multidimensionality using PCA in order to aggregate risk and forecast applications. SVM works best with the model because of its 82% accuracy.

TABLE II
AUTHORS AND METHODOLOGIES USED

Authors	Methodologies Used
E S, Park J, Cho Y	Linear regression, Gradient boosting, and SVM
Ceccaeelli G and Cho Y	Random forests, Gradient boosting
Jiang	CNN, RNN, and Graph neural networks
Jelic A. and Roncagila P.	Linear regression, Ridge and Lasso regression, SVM, Decision Tree, Random Forest regressor, and ARIMA
Collini E, Nesi P, and Pantaleo G	LSTM, GRU, XGBoost, DNN, and Random Forest

Nowadays, rental bikes are being offered in several places to improve the ease of mobility. [6] have used models like Linear Regression, Gradient Boosting, and Support Vector Machine (SVM) with optimum hyperparameters utilizing repeated cross-validation. In order to improve mobility comfort, rental bikes are currently available in many urban areas. Gradient Boosting emerged as the most effective model, yielding the highest and best R^2 values (0.92) in the testing set and 0.96 in the train set when all predictors are applied.

In a separate study, Ceccaeelli G and Cho Y [7] created machine learning classifiers, mainly using random forests and gradient boosting, to forecast inventory levels based on historical data, with a particular emphasis on stations. To deal with the imbalanced data, they have also employed the SMOTE technique.

Author Jiang [8] employed a deep learning model to forecast the bike inventory. There was use of DL models, including CNN, RNN, and Graph neural networks. The ARIMA time series model is employed when temporal data that must be predicted over time is present. The state of the art was determined to be RNNs, and the state-of-the-art solutions were GCN and GAT.

Models such as Linear regression, Ridge and Lasso regression, SVM, Decision Tree, Random forest regressor, and ARIMA have also been used by Jelic A. and Roncagila P. [9]. ARIMA with 0.19 RMSE was the best fit model that was appropriate and provided better accuracy, about 92%.

The quantity of bikes that are available and the number of open bike slots in bike-sharing stations have been predicted by Collini E, Nesi P, and Pantaleo G [10] using a number of models, including LSTM, GRU, XGBoost, DNN, and Random Forest, to make this prediction. Short-term forecasts for the upcoming 15, 30, 45, and 60 minutes were made using the high-dimensional time series, which included forecasts for each station and the weather.

The Author Ignatius Hermina C [11] has implemented MBA in order to understand consumer purchasing trends and recommend items based on their previous purchase history. They have used Apriori and FP-growth to improve precision. FP-growth has proven significantly quicker compared to Apriori, resulting in the identification of mineral water as the most purchased item.

Rao A and Kiran J [12] have applied MBA in the healthcare domain, helping residents become aware of frequently upcoming diseases to take timely actions. They also utilized Apriori to make recommendations using association rules.

Apriori algorithm is employed by Sai Shree V and Lakshmi A [14] to increase sales, ensuring smooth and profitable business operations for the organization. The complexity was reduced through the execution and implementation of hash tables.

III. Methodology

Knowledge Discovery in Databases (KDD) is a comprehensive approach to data mining that involves taking large datasets and extracting relevant patterns, data insights, and knowledge. The key components of the KDD process, encompassing data selection, preprocessing, transformation, mining, evaluation, and interpretation which can be seen in Fig.1 [11].

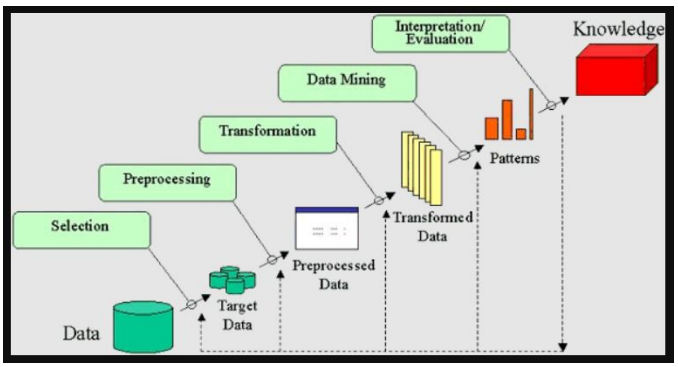

Fig. 1. Flow of knowledge discovery in databases (KDD) technique [14]

Emphasis is placed on the iterative nature of KDD, where feedback loops refine the process and enhance its efficacy. A critical aspect of KDD is data preprocessing, where raw data undergoes cleansing and transformation to ensure its quality and relevance. Various techniques such as handling missing values, outlier detection, and normalization are discussed in detail, underscoring their role in preparing data for meaningful analysis. The mining phase, a central component of KDD, involves the application of advanced algorithms to extract patterns and knowledge from the pre-processed data. Classification, clustering, association rule mining, and anomaly detection are explored as fundamental techniques contributing to the diverse set of insights that KDD can reveal. The evaluation of discovered knowledge, elucidating the importance of assessing the quality and reliability of the extracted patterns. Metrics such as accuracy, precision, recall, and F1-score are presented as essential tools for evaluating the performance of data mining models. Challenges and ethical considerations related to KDD are also addressed, acknowledging issues such as privacy concerns, biased data, and the interpretability of machine learning models. Data Mining emphasizes the need for responsible and ethical practices in KDD to ensure the fair and transparent use of data-driven insights.

A. Data Selection

The first step is to establish and understand the dataset in order to satisfy the end user's desire to determine whether the dataset is relevant. If the quality of the existing data is low, there is a need to collect new raw data based on the goal. Domain knowledge is always required in order to comprehend the data.

1) Dataset 1: The dataset has been acquired from UCI Machine learning repository. It has 25 attributes with 30000 records based on Taiwan credit card users as a defaulter. Binary classification values, such as 0 and1, which denote No and Yes, respectively, make up the target variable. Our analysis will be forecast credit card defaults in the future based on individuals account information and past bill payment history. The descriptive statistics such as measuring central tendency like mean, median, model and the percentile details are shown in the Fig.

	count	mean	std	min	25%	50%	75%	max
LIMIT_BAL	30000.0	167484.322667	129747.661567	10000.0	50000.00	140000.0	240000.00	1000000.0
AGE	30000.0	35.485500	9.217904	21.0	28.00	34.0	41.00	79.0
PAY_SEPT	30000.0	-0.016700	1.123802	-2.0	-1.00	0.0	0.00	8.0
PAY_AUG	30000.0	-0.133767	1.197186	-2.0	-1.00	0.0	0.00	8.0
PAY_JUL	30000.0	-0.166200	1.196868	-2.0	-1.00	0.0	0.00	8.0
PAY_JUN	30000.0	-0.220667	1.169139	-2.0	-1.00	0.0	0.00	8.0
PAY_MAY	30000.0	-0.266200	1.133187	-2.0	-1.00	0.0	0.00	8.0
PAY_APR	30000.0	-0.291100	1.149988	-2.0	-1.00	0.0	0.00	8.0
BILL_AMT_SEPT	30000.0	51223.330900	73635.860576	-165580.0	3558.75	22381.5	67091.00	964511.0
BILL_AMT_AUG	30000.0	49179.075167	71173.768783	-69777.0	2984.75	21200.0	64006.25	983931.0
BILL_AMT_JUL	30000.0	47013.154800	69349.387427	-157264.0	2666.25	20088.5	60164.75	1664089.0
BILL_AMT_JUN	30000.0	43262.948967	64332.856134	-170000.0	2326.75	19052.0	54506.00	891586.0
BILL_AMT_MAY	30000.0	40311.400967	60797.155770	-81334.0	1763.00	18104.5	50190.50	927171.0
BILL_AMT_APR	30000.0	38871.760400	59554.107537	-339603.0	1256.00	17071.0	49198.25	961664.0
PAY_AMT_SEPT	30000.0	5663.580500	16563.280354	0.0	1000.00	2100.0	5006.00	873552.0
PAY_AMT_AUG	30000.0	5921.163500	23040.870402	0.0	833.00	2009.0	5000.00	1684259.0
PAY_AMT_JUL	30000.0	5225.681500	17606.961470	0.0	390.00	1800.0	4505.00	896040.0
PAY_AMT_JUN	30000.0	4826.076867	15666.159744	0.0	296.00	1500.0	4013.25	621000.0
PAY_AMT_MAY	30000.0	4799.387633	15278.305679	0.0	252.50	1500.0	4031.50	426529.0
PAY_AMT_APR	30000.0	5215.502567	17777.465775	0.0	117.75	1500.0	4000.00	528666.0

Fig. 2. Descriptive statistics for Credit card Default (Dataset 1)

2) Dataset 2: This dataset, comprising 17379 records and 17 attributes, was also acquired from the UCI repository. The target variable is the numerical total counts of bikes rented. The analysis needs to be done in order to predict the upcoming year's count of rental bikes. The statistical measures for this dataset can also be seen in Fig.

	count	mean	std	min	25%	50%	75%	max
instant	17379.0	8690.000000	5017.029500	1.00	4345.5000	8690.0000	13034.5000	17379.0000
season	17379.0	2.501640	1.106918	1.00	2.0000	3.0000	3.0000	4.0000
yr	17379.0	0.502561	0.500008	0.00	0.0000	1.0000	1.0000	1.0000
mnth	17379.0	6.537775	3.438776	1.00	4.0000	7.0000	10.0000	12.0000
hr	17379.0	11.546752	6.914405	0.00	6.0000	12.0000	18.0000	23.0000
holiday	17379.0	0.028770	0.167165	0.00	0.0000	0.0000	0.0000	1.0000
weekday	17379.0	3.003683	2.005771	0.00	1.0000	3.0000	5.0000	6.0000
workingday	17379.0	0.682721	0.465431	0.00	0.0000	1.0000	1.0000	1.0000
weathersit	17379.0	1.425283	0.639357	1.00	1.0000	1.0000	2.0000	4.0000
temp	17379.0	0.496987	0.192556	0.02	0.3400	0.5000	0.6600	1.0000
atemp	17379.0	0.475775	0.171850	0.00	0.3333	0.4848	0.6212	1.0000
hum	17379.0	0.627229	0.192930	0.00	0.4800	0.6300	0.7800	1.0000
windspeed	17379.0	0.190098	0.122340	0.00	0.1045	0.1940	0.2537	0.8507
casual	17379.0	35.676218	49.305030	0.00	4.0000	17.0000	48.0000	367.0000
registered	17379.0	153.786869	151.357286	0.00	34.0000	115.0000	220.0000	886.0000
cnt	17379.0	189.463088	181.387599	1.00	40.0000	142.0000	281.0000	977.0000

Fig. 3. Descriptive statistics for Bike sharing demand prediction (Dataset 2)

3) Dataset 3: This dataset has been taken from the retail store of United Kingdom which is in CSV format. The data

consist of information about the stores, products and the transactions from the store. The product file has data that have been sold which includes European Article Number (EAN) which is unique, short and long description, product and price. The transaction data includes features such as Basket ID, total number of products in the transaction, Request basket JSON comprising basket data and the time of the transaction. The quality of data has been generated in order to know the data.

B. Data Preprocessing and Transformation

Following data selection, the next stage is the data preprocessing stage. In this stage, the raw data undergoes cleansing, transformation, and integration processes. Data needs to be cleaned by handling missing values, outliers, and errors, ensuring the data is reliable and accurate. Transformation such as standardization and normalization is applied to maintain data consistency. In order to maintain the complexity of the dataset, techniques such as feature selection and dimensionality reduction are used to streamline the data.

1) Dataset1: There are no null or duplicated values present in the data. To enhance interpretability and gain valuable insights, mapping procedures have been implemented for several key columns, namely sex, education, marriage, and IsDefaulter. The explanatory data analysis has been performed and analytical approach encompasses a spectrum of techniques, including univariate, bivariate, and multivariate analyses. From

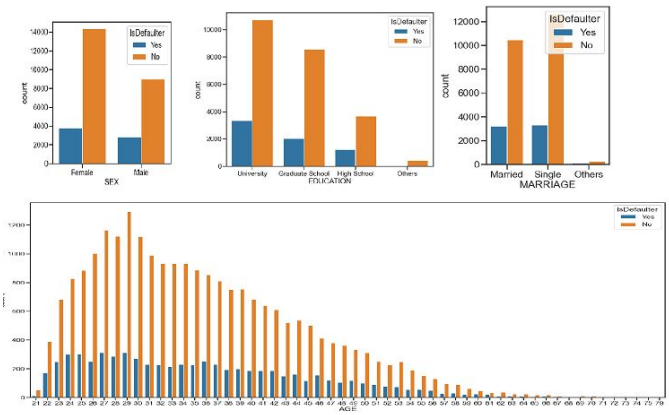

Fig. 4. Bivariate analysis of credit card dataset

the Fig 4, the visualization suggests a correlation between gender and credit default, with a higher proportion of defaulters identified as female. Additionally, the examination of education levels reveals that individuals classified as university and graduate students are more prone to default on credit. Additionally, a significant percentage of the defaulters in this subset identify as single. The variable 'PAY' is highly correlated with the others, while there is only a moderate positive correlation seen in Fig. . Every single one of the data set's "BILL AMT" variables has a strong positive correlation with one another as well. The outliers have been detected and treated using z-score in order to get better accuracy. As the values of all the features vary from different ranges, the author

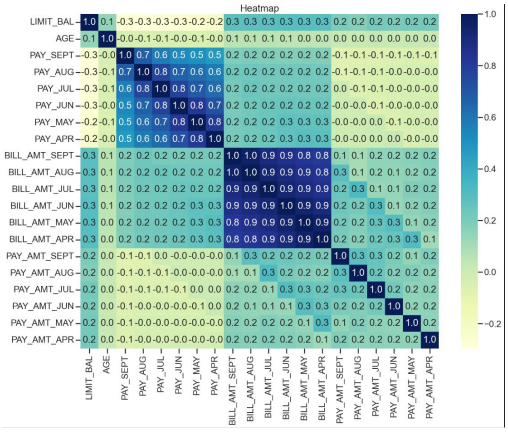

Fig. 5. Heatmap of credit card default

has performed the standardization using Min-Max scaler. The Min-Max scaler operates by rescaling the range of values within each feature to a predefined interval, typically between 0 and 1.

2) Dataset 2: No null values are present and all the records are unique. We have extract day, month and year from the date column in order to visualize in a better way. On this dataset. EDA was conducted using univariate, bivariate and multivariate analysis. Fig.6.In delving into the univariate aspects,

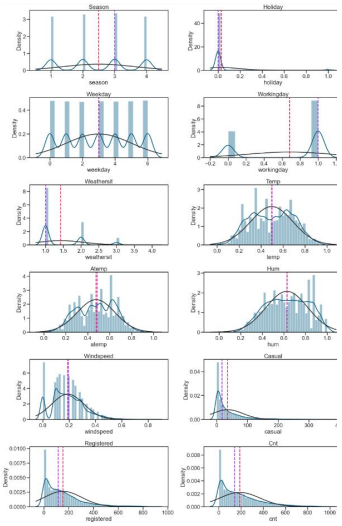

Fig. 6. Univariate Analysis of Bike sharing Demand

distinct distributions of various features have been unveiled. Notably, the features associated with holidays, weather situations, windspeed, casual riders, registered riders, and overall count exhibit positive skewed distributions. This skewness suggests a concentration of values towards the lower end of the distribution, indicating potential patterns or trends in the data. This absence of skewness suggests a more symmetric distribution of values, highlighting a balanced representation across the range. Furthermore, an intriguing observation emerges in the analysis of features such as working day, season, and

weekday, which portray negative skewed distributions. This skewness implies a concentration of values towards the higher end, underscoring potential trends in these variables. Fig, The

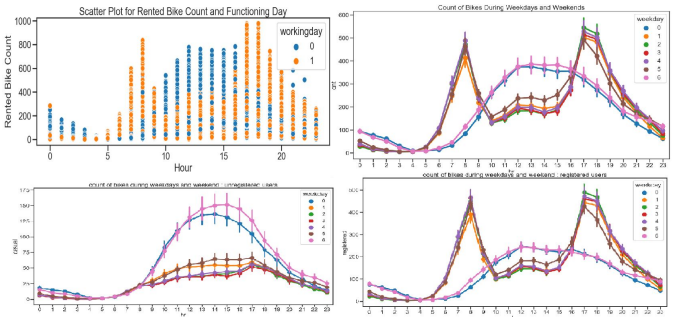

Fig. 7. Bivariate Analysing of bike sharing dataset

visual representation distinctly reveals heightened demand for rental bikes on weekdays during peak hours (5 am-8 am and 4 pm-7 pm), reflecting commuting patterns to and from work. On weekends, a consistent demand spans the entire day, particularly in the afternoon, indicative of a more leisurely and varied usage pattern during non-working days. To preserve the integrity of the dataset, outliers in the 'casual' and 'registered' features were deliberately disregarded, considering the 'cnt' variable is their sum. Treating these outliers could potentially distort the overall distribution, leading to inaccurate results. Employing label encoding was essential to convert categorical data into a numerical format, prioritizing rank as a crucial factor. Further enhancing the data preprocessing, Min-Max scaling was applied, ensuring a standardized scale across all features. The target variable 'cnt' exhibited positive skewness, prompting the exploration of log transformation to rectify this skewness and potentially improve model performance. These meticulous steps in data preprocessing lay the foundation for a more reliable and robust analysis, fostering accurate insights and predictions in subsequent modeling endeavors.

3) Dataset 3: Initially the semi-structured data was converted into structured data which creates the one record for each of the product. Example if the transaction contains 5 items, then that number of rows will get created. The missing and inconsistent product ids which we removed in order to have the clean data. The column has the temporal variable called timestamp based on it the new columns has been created which is an categorical variable classified into morning, afternoon, evening and night. The frequency has been calculated for each and every product and has been stored in a new column. As the cost keeps changing, therefore new column has been introduced known as 'Effective price' which consist of average cost of product. As we will not be able to recommend having only one item as a transaction, hence such records have been removed. As the data were grouped in order to convert it into structured data, it has been grouped back based on transaction ID.

C. Data Mining

The heart of data mining lies in the application of algorithms and statistical models to identifying patterns, associations, classifications, and anomalies in the data. In this process, ML is essential as it allows systems to learn from dataset and make decisions or predictions. Applications of data mining are widespread and impact areas such as business, finance, healthcare, marketing, and more. In business, it aids in customer segmentation, market basket analysis, and fraud detection. In the analysis of the dataset, a repertoire of machine learning algorithms , namely Logistic Regression, Decision Tree, Random Forest , and Support Vector Machine (SVM), has been deployed. This diverse ensemble offers a comprehensive approach to classification, enabling the prediction of whether a consumer is likely to default on credit. To optimize these algorithms and enhance their performance, hyperparameter tuning has been implemented, ensuring a more efficient and accurate classification of new data. Logistic Regression emerges as a particularly pertinent choice in this analysis. It excels in both classification and regression tasks, making it versatile for predicting credit card defaults. The algorithm's equation, akin to linear regression, is adept at handling binary outcomes, aligning seamlessly with the nature of the dataset where the goal is to classify clients into default or non-default categories. The simplicity and interpretability of Logistic Regression contribute to its utility, providing insights into the factors influencing credit default. In contrast, Decision Trees, Random Forests, and SVM each bring unique strengths to the predictive modelling landscape. Decision Trees are advantageous for datasets with numerous categorical variables, as they efficiently partition the data based on these variables. However, their interpretability is limited, and they may suffer from overfitting. Random Forests, an ensemble of Decision Trees, overcome some of these limitations by aggregating predictions, offering improved accuracy and generalization. SVM, on the other hand, are adept at handling complex datasets through the creation of hyperplanes that best separate different classes. Despite being less interpretable and slower in producing predictions compared to a Decision Tree, the Support Vector Machine excels in scenarios where intricate decision boundaries exist. The intricate nature of credit default prediction, influenced by various factors, makes SVM a valuable contender. The rationale behind choosing a specific algorithm often hinges on a trade-off between interpretability and predictive performance. In this context, the selection of Random Forests indicates a strategic choice. While less interpretable due to the complexity of multiple decision trees, the model outperforms a standalone Decision Tree, offering superior predictive accuracy. The focus here is on achieving the best possible performance and accurately predicting credit default, justifying the preference for a more sophisticated model like Random Forest. Apriori is an algorithm based on Association rule mining (ARM) to revile the uncover frequent pattern. It is widely used for capturing the patterns, associations for market basket data and the relationships.

The main intention is to identify the set of items frequently occurring together and the hiding association. FP Growth plays a significant role order to build a compact data refereed as FP tree. It has 2 phases, the phase one, it constructs the FP tree by identifying frequent items and the phase two consist of mining those frequent patterns. It applies the divide and conquer strategy, resulting into breaking down the mining into chunks, which reduces the database scans. As the dataset is imbalanced, so we have used the SMOTE technique. SMOTE helps to make the data balance by over sampling it.

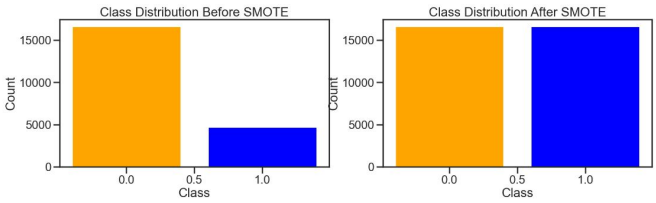

Fig. 8. SMOTE technique to treat the imbalanced data

IV. Evaluation

In the evaluation phase of the Knowledge Discovery in Databases (KDD) process, the effectiveness of the data mining algorithms employed is rigorously assessed. This involves scrutinizing the results and outputs to determine the presence of acceptable and useful patterns. The evaluation encompasses various parameters, offering a comprehensive assessment of the models developed during the data mining phase.

A. Dataset 1

Initially, logistic regression was created by obtaining 81% precision, 80% accuracy, and high recall at 97%. All these were maintained with an AUC of 73%, even after hyper-parameter tuning using `GridSearchCV`. Once the `SMOTE` technique, used to treat imbalanced data, was applied, the accuracy dropped to 66%, recall to 64%, and precision to 67%. However, after another round of hyperparameter tuning, precision rose to 87%. These emphasize how dynamic model performance is and how the algorithm makes decisions. Next, the Decision Tree was applied, demonstrating the performance of a binary classification model. Achieving 83% precision and 95% recall for class '0', with an F1 score of 89%. For class '1' (individuals with defaulters), precision was 68% and recall was 36%, resulting in an F1-score of 47%. The overall F1 score was 79%, and accuracy was 81%. After applying the `SMOTE` technique, a trade-off between precision and recall was observed. Precision decreased to 27%, while recall increased to 88% for class '0'. The model's accuracy dropped to 44%, indicating challenges in predicting both classes accurately. The macro average F1-score was 40%, showing a balanced trade-off between recall and precision. The weighted average F1-score and accuracy were 39% and 40%, respectively, indicating difficulties in achieving balanced prediction outcomes. The Random Forest model showed 83% precision with 96% recall for class '0', resulting in an F1 score of 89%. For class '1', the F1-score was 47% due to precision and recall at 70% and 35%, respectively. Even after hyperparameter tuning, the model's metrics remained substantial with 83% precision and 95% recall for class '0'. The Random Forest model with GridSearchCV demonstrated good predictive capabilities, achieving a high precision and overall accuracy of 90%.

Logistic Regression (GridSearchCV)

	precision	recall	f1-score	support
0	0.81	0.97	0.88	4117
1	0.69	0.23	0.35	1223
accuracy			0.80	5340
macro avg	0.75	0.60	0.61	5340
weighted avg	0.78	0.80	0.76	5340

Logistic Regression (with SMOTE)

	precision	recall	f1-score	support
0	0.87	0.65	0.75	4117
1	0.37	0.67	0.47	1223
accuracy			0.66	5340
macro avg	0.62	0.66	0.61	5340
weighted avg	0.76	0.66	0.68	5340

Fig. 9. Classification report for Logistic regression

Decision Tree using GridSearchCV

	precision	recall	f1-score	support
0	0.83	0.96	0.89	4117
1	0.70	0.35	0.47	1223
accuracy			0.82	5340
macro avg	0.76	0.65	0.68	5340
weighted avg	0.80	0.82	0.79	5340

Decision Tree using GridSearchCV with SMOTE

	precision	recall	f1-score	support
0	0.90	0.25	0.39	4117
1	0.26	0.91	0.41	1223
accuracy			0.40	5340
macro avg	0.58	0.58	0.40	5340
weighted avg	0.75	0.40	0.39	5340

Fig. 10. Classification report for Decision Tree

B. Dataset 2

For the bike sharing demand data, KNN and Linear regression models have been used.

For Linear regression, R^2 score is 0.734 which indicates that around 73.4% of the variance in the target variable.

Random Forest using GridSearchCV

	precision	recall	f1-score	support
0	0.83	0.95	0.89	4117
1	0.67	0.35	0.46	1223
accuracy			0.81	5340
macro avg	0.75	0.65	0.67	5340
weighted avg	0.79	0.81	0.79	5340

Random Forest using Gridsearch CV with SMOTE

	precision	recall	f1-score	support
0	0.83	0.96	0.89	4117
1	0.70	0.35	0.47	1223
accuracy			0.82	5340
macro avg	0.76	0.65	0.68	5340
weighted avg	0.80	0.82	0.79	5340

Fig. 11. Classification report for Random Forest

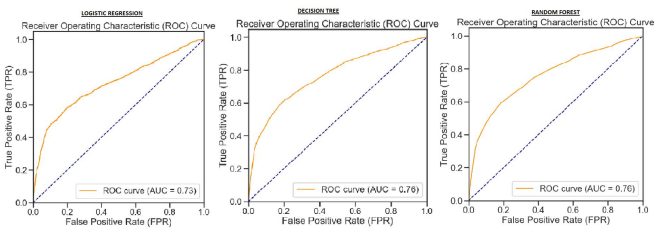

Fig. 12. ROC-AUC curve for dataset 1

The adjusted R^2 for the training set is 73% which provides the measure that adjusts for the number of predictors. The MSE is 0.545, representing the average square difference between actual and predicted with an RMSE of 0.739. In contrast, the KNN demonstrates R^2 score of 0.873, indicating a better fit compared to Linear regression. The training set's Adjusted R^2 of 0.873 indicates how well the model captured the variability in the target variable. The MSE and RMSE for the model are 0.260 and 0.510, respectively. Overall, the KNeighborsRegressor performs better than Linear Regression, as evidenced by its higher R^2 scores, lower MSE, and RMSE values, which suggest that it is more successful at identifying the underlying patterns in the data.

C. Dataset 3

The MBA have bene used in order to recommend the items based on ARM. The training involves utilising the Apriori which extracts the item and rules from the transaction. Which are further filtered into the significant association between items. Th training process handles empty values and rounding numerical attributes. The model predictions recommend the items that were extracted and converted into JSON. Overall, the MBA generates the meaningful recommendation based on ARM deriving from transactional data.

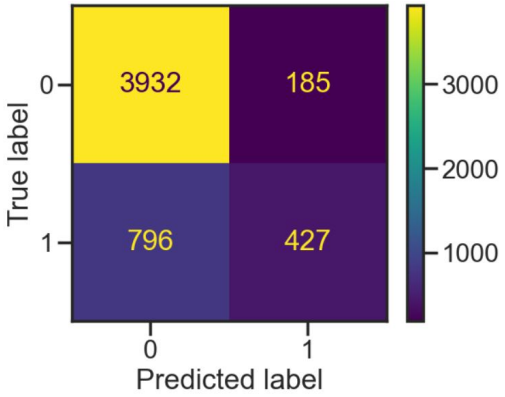

Fig. 13. Confusion Matrix for Dataset 1

	Model	R2_score	Adjusted_R2_train	MSE	RMSE
0	Linear_Regression	0.734205	0.733452	0.545395	0.738508
1	KNeighborsRegressor	0.873074	0.872715	0.260444	0.510337

Fig. 14. Summary for Dataset 2

V. Conclusion

For **Dataset 1**, the logistic regression initially performed well with a precision of 81%, accuracy of 80%, and a high recall of 97%. Although these metrics, with an AUC of 73%, improved through hyperparameter tuning using GridSearchCV, the application of the SMOTE process after balancing the data led to a drop in accuracy. However, subsequent parameter tuning resulted in a rise in precision. On the other hand, the decision tree in Dataset 1 showed a trade-off between precision and recall, especially after applying SMOTE. The balanced metrics performed poorly, with an overall accuracy of 44%, indicating challenges in achieving accurate predictions for both classes. In contrast, the random forest performed exceptionally well, showcasing metrics with 83% precision and 96% recall, achieving an overall accuracy of 90% after optimizing using hyperparameter tuning. This model worked extremely well in handling imbalanced data. For **Dataset 2** on bike-sharing demand, linear regression and KNN were performed. The KNN outperformed with high R^2 scores, lower MSE, and RMSE values, indicating better success in identifying patterns. The demand for rental bike grows of registerd users during working day, while for unregistered users it grows during the weekend as the indiviaduals go out on a trip during the weekend. **Dataset 3** was used for implementing Market Basket Analysis, successfully leveraging the Apriori Algorithm to extract association rules. This approach offers a scalable solution for businesses to understand item leverage, recommend items, and improve decision-making.

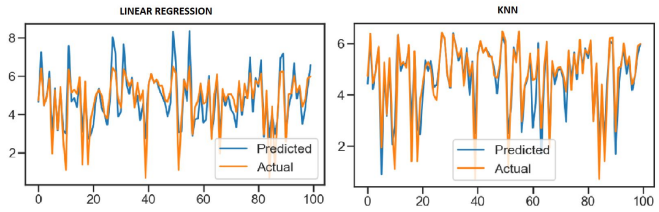

Fig. 15. Actual vs Predicted value for Linear regression and KNN

	RecommendedItem
1	5000128271165
2	5000128871457
0	5010891011868
0	5010891011868

Fig. 16. Recommended items

REFERENCES

[1] Y. Chen and R. Zhang, "Research on Credit Card Default Prediction Based on k-Means SMOTE and BP Neural Network," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/6618841.

[2] R. Liu and R. Liu, "Machine Learning Approaches to Predict Default of Credit Card Clients," *Modern Economy*, vol. 9, no. 11, pp. 1828–1838, Nov. 2018, doi: 10.4236/ME.2018.911115.

[3] Y. Ma, "Prediction of Default Probability of Credit-Card Bills," Open Journal of Business and Management, vol. 08, no. 01, pp. 231–244, 2020, doi: 10.4236/ojbm.2020.81014.

[4] A. Bacova and F. Babic, "Predictive Analytics for Default of Credit Card Clients," SAMI 2021 - IEEE 19th World Symposium on Applied Machine Intelligence and Informatics, Proceedings, pp. 329–333, Jan. 2021, doi: 10.1109/SAMI50585.2021.9378671.

[5] S. Arora, S. Bindra, S. Singh, and V. Kumar Nassa, "Prediction of credit card defaults through data analysis and machine learning techniques," Mater Today Proc, vol. 51, pp. 110–117, Jan. 2022, doi: 10.1016/J.MATPR.2021.04.588.

[6] S. V. E, J. Park, and Y. Cho, "Using data mining techniques for bike sharing demand prediction in metropolitan city," Comput Commun, vol. 153, pp. 353–366, Mar. 2020, doi: 10.1016/J.COMCOM.2020.02.007.

[7] G. Ceccarelli, G. Cantelmo, M. Nigro, and C. Antoniou, "Learning from Imbalanced Datasets: The Bike-Sharing Inventory Problem Using Sparse Information †," Algorithms, vol. 16, no. 7, Jul. 2023, doi: 10.3390/a16070351.

[8] W. Jiang, "Bike sharing usage prediction with deep learning: a survey," Neural Computing and Applications. Springer Science and Business Media Deutschland GmbH, 2022. doi: 10.1007/s00521-022-07380-5.

[9] A. Jelic and P. Roncaglia, "Predicting bike sharing demand with machine learning," 2021.

[10] E. Collini, P. Nesi, and G. Pantaleo, "Deep Learning for Short-Term Prediction of Available Bikes on Bike-Sharing Stations," IEEE Access, vol. 9, pp. 124337–124347, 2021, doi: 10.1109/ACCESS.2021.3110794.

[11] C. Ignatius Hermina, "MARKET BASKET ANALYSIS FOR A SUPERMARKET." [Online]. Available: https://www.researchgate.net/publication/365489098

[12] A. B. Rao, J. S. Kiran, and P. G, "Application of market–basket analysis on healthcare," International Journal of Systems Assurance Engineering and Management, Dec. 2021, doi: 10.1007/s13198-021-01298-2.

[13] Vj. Sai Sree, Aa. Lakshmi, and J. Jayanthi, "Market Basket Analysis using Apriori Algorithm," 2022.

[14] "KDD Process/Overview." Accessed: Jan. 04, 2024. [Online]. Available: (https://www2.cs.uregina.ca/ dbd/cs831/notes/kdd/1$_k dd.html$