Long Term Deposit Prediction of Bank Using Data Mining Techniques

Jitendhar Reddy Adulla, Joshna Devi Vadapalli, Shrivatson Ramaratnam Giridharan, Syama Ravi Teja Jerrypothula, Vamshi Krushna Lakavath

Department of Applied Data Science

San Jose State University

Data 240: Data Mining and Analytics

Dr. Seungjoon (Joon) Lee

Dec 13, 2022

Long-Term Deposit Prediction of Bank Using Data Mining Techniques

Introduction

In today's world, one of the best ways an organization can improve its performance in the market is by capturing and analyzing the customer data in an efficient way for better customer experience. And a huge amount of data is getting generated in day-to-day activities of various fields, where the banking sector is one of them. In recent times the finance industry has transformed the most by machine learning technology. There are two main approaches for industries in marketing such as mass campaigns where they target the public randomly and the other is direct marketing where the target is a set of users. Nowadays the response to the mass campaigns is minimal compared to direct marketing, many banking institutions prefer to use direct campaigns via phone calls which showed significant success rate.

With the help of technology they can make better marketing strategies. Since data mining technology is becoming more mature and widely used in customer-oriented businesses and banking institutions to predict target customer groups for better sales. Therefore a data driven marketing strategy is required which relies on data analysis to identify patterns and trends in customer behavior that helps banks identify the target customer group better. Therefore, for our project we choose to analyze the dataset taken from UCI machine learning repository which is related to banking campaigns in Portuguese banks via direct phone calls. Based on the bank marketing data, we are going to propose a model to predict whether the customer has opted for a term deposit or not. In this project, Python is used as a coding language and machine learning algorithms for statistical analysis purposes. Since our data is labeled data we are planning to use supervised machine learning models such as Random Forest, Support Vector Machine, K-nearest neighbors, and Naïve Bayes for classification and analysis. The performance of the results are evaluated based on the metrics, R2, RMSE, F1 score, accuracy, precision to decide on the better model for prediction. These classification results help banks group various customers like high-valued customers, potential customers to improve marketing efficiency.

For this problem we collected Bank Marketing data from Data World website (2017). This dataset consists of 21 columns including the target column and with more than 40K rows. The description of this dataset is described in Table 1. As mentioned in that table, many of the columns are categorical. The possible categories of those columns are clearly mentioned in Table 2.

Table 1

Metadata of Data Set

S. No.	Column Name	Data Type	Column Description
1.	age	Numeric	Age of the bank customer.
2.	job	Categorical	Customers' job type.
3.	martial	Categorical	marital status of the customer
4.	education	Categorical	Highest education of a customer

5.	default	Categorical	Whether the loan taken by customer is default or not?
6.	housing	Categorical	Whether the customer has availed housing loan?
7.	loan	Categorical	If the customer has personal loan?
8.	contact	Categorical	What kind of contact communication type customer has?
9.	month	Categorical	In which month of the year a customer was last contacted
10.	day_of_week	Categorical	Last contact day of the week with the customer
11.	duration	Numeric	last contact duration with customer, in seconds
12.	campaign	Numeric	number of times contacts performed during this campaign and for this client, including last contact.
13.	pdays	Numeric	number of days that passed by after the customer was last contacted from a previous campaign
14.	previous	Numeric	number of contacts performed before this campaign and for each customer
15.	poutcome	Categorical	result of the previous marketing campaign Like
16.	emp.var.rate	Numeric	employment variation rate - quarterly indicator
17.	cons.price.idx	Numeric	consumer price index - monthly indicator
18.	cons.conf.idx	Numeric	consumer confidence index - monthly indicator
19.	euribor3m	Numeric	euribor 3 month rate - daily indicator
20.	nr.employed	Numeric	number of employees - quarterly indicator
21.	y	Binary	Whether the client subscribed a term deposit or not?

Table 2

Possible Categorical Values of Columns

S. No.	Column Name	Possible Categorical Values
1.	job	'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self employed', 'services', 'student', 'technician', 'unemployed', 'unknown'

2.	martial	'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed
3.	education	'basic.4y' ,'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown'
4.	default	'no', 'yes', or 'unknown'
5.	housing	'no', 'yes', or 'unknown'
6.	loan	'no', 'yes', or 'unknown'
7.	contact	'cellular', 'telephone'
8.	month	'jan', 'feb', 'mar', ..., 'nov', 'dec'
9.	day_of_week	'mon', 'tue', 'wed', 'thu','fri'
10.	poutcome	'failure', 'nonexistent', 'success'
11.	y	'yes', 'no'

Need of Classification Method

A weather assessment system should have the ability to classify the chance of rain during a particular day and should be able to forecast the same for a stipulated amount of days. A bank employee needs to answer the question about the person's credibility before issuing a loan or credit card. So, usually before making some decisions you need some answers for the questions that constitute the decision. For these kinds of situations classification methods are the best suitable techniques. Classification methods will classify the given data into one of the output classes like raining or not raining for the aforementioned example. Classification utilizes the training dataset to improve the boundary conditions that will be used to identify each target class. After determining the boundary conditions, the next step is to forecast the target class. We have many classification techniques like Decision Trees (DTs), Support Vector Machine (SVM), Logistic Regression, Naive Bayes etc.

Apart from these, some special models like ensemble learning which combine multiple models to improve performance of models. In our project we need an answer for the question of whether the customer has opted for long-term deposit or not. We considered a dataset of a Portuguese bank which consists of attributes of questions about marriage, education, housing, loan and etc with the target attribute long-term deposit. The whole dataset is classified into two groups i.e., opted for long-term deposit or not. Therefore classifications techniques are the best options to deal with this problem. These methods will help us to come up with a boundary between those two classes and also helpful to analyze the importance of input variables/attributes on the target class.

Literature Review

This paper is about the bank marketing successful estimation using data mining techniques. Author considered a real dataset from a Portuguese retail bank which consists of 52K records with 150 features. As there are more features there is a necessity of feature extraction. In this paper, the author used a semi-automatic approach for feature extraction which involves two steps. The first step is they used domain knowledge and made 14 questions to filter features. After that they used wrapper techniques to extract important features. After applying all these, they came up with 22 features (Moro et al., 2014).

Models were developed using four different models in data mining. Those are Decision Trees, Support Vector Machine (SVM), Logistic Regression and Neural Network (NN). For evaluating these models, authors used Area of the Receiver Operating Characteristic Curve (AUC) and Area of the LIFT Cumulative Curve (ALIFT). Among all the developed models, NN outperformed all other models with AUC = 0.8 and ALIFT = 0.7. So, this model is useful to banks to predict the long term subscription of the customer before doing a telephone call itself. Which will be easy to filter customers and target them. Apart from that, from this model they can also know which features are affecting a lot of decision making of long term deposits by customers (Moro et al., 2014).

Çığşar & Ünal stated that recent years have seen a sharp rise in the use of big data and its analysis, particularly in the banking and finance sector. They stated that Data Mining is a statistically driven data analysis technique that aims to uncover knowledge from vast amounts of data. In this research, they have conducted a comparative analysis of data mining and classification algorithms to determine the risk of a bank account becoming the default. They have used the credit card dataset which was collected from the TUIK survey in 2015, which contains 12 fields and one of them is a class variable payment or non-payment (Çığşar & Ünal, 2019).

The main objective of the classification method is to correctly forecast the target class of objects for which the class label is unknown. For their study, they have built Bayes networks (BayesNet), Naive Bayes, logistic regression, multilayer perceptron; J48, and Random forest algorithms using the WEKA software. Accuracy, Precision, Recall, F-measure, ROC area, and RMSE evaluation metrics have been used, and determined Logistic Regression as the best model for classifying credit card accounts with 83.1% of accuracy. They concluded that the risk can be mitigated by using the data mining classification methods to determine the class of a credit account and take precautionary actions. That can lower the cost and losses, and increase the profits of the financial organizations (Çığşar & Ünal, 2019).

In this paper, Wang(2020) states that datamining is used prominently in the field of banking and finance to identify target customer groups and thus improve bank sales. He uses the bank dataset from UCI machine learning repository which is based on telephone tracking of Portuguese banks (Wang, 2020).

The dataset has 4521 data with 16 independent variables and 1 dependent variable. Initially the dataset has been analyzed to achieve the quality of data by replacing the outliers or extreme values with normal data. He uses optimal feature variable selection to avoid redundant variables. Also by classifying and constructing a decision tree for all the 16 variables, he found the strongly correlated variables to choose final characteristic variables among them. In order to do better classification he has used a random sampling method to process the unbalanced data. In this article he uses the C5.0 decision tree model which generates decision tree model and rule set model. In the decision tree each leaf node represents a subset of data. Rule set

simplifies the description of the information in the decision tree model. Based on the rules used in decision tree, he found various patterns which are useful for banking purposes (Wang, 2020).

To evaluate the performance various evaluation metrics have been taken such as accuracy, specificity, sensitivity and confusion matrix to better visualize the performance. The dataset is divided into 60% train set and 40% test set. The results are then compared with other four classification algorithms: C&T, CHAID, Neural networks, QUEST. Based on the results he concluded that the C5.0 and C&T model gives better prediction of the target customer group classification. The obtained classification rules help banks improve sales and make marketing better (Wang, 2020).

Alexandra & Sinaga (2021) showed the importance of decision support systems which plays a crucial role in the business planning and outcomes. Banking sector being a financial institution needs to have a very firm decision making inorder to make an impact in the space. Bank product marketing should conduct market research and analysis to provide potential customers with the best products. They applied various machine learning techniques to real data from Portuguese banks and found it challenging to predict the behavior of the customers. Machine learning techniques such as Naive Bayes, Random Forest and various unsupervised learning techniques such as K-Means, DBSCAN clustering were employed. A Portuguese dataset containing 45211 marketing data with 17 attributes was used. They contained features such as age, income, marital status, loan information etc. Performance evaluation metrics such as Balanced Accuracy (BA) and Mathews Correlation Coefficient(MCA) were used. Since this dataset can be used for both supervised and unsupervised problems, machine learning techniques such as both classification and clustering were deployed.

Since the data labels have been defined already classification techniques such as logistic regression were deployed on the data. The data for churn prediction was labeled as 0 for not churn and 1 for churn. This takes into account the cost function which is the difference between predicted label and the ground truth. Thus smaller the values of the cost function the model is performing better. Decision tree is one which does not worry about outliers since it splits the data based on various criteria. Random Forest is one of the best ensemble methods since it uses majority voting and it splits the data into various chunks named as bootstrap samples. Naive Bayes uses Bayes theorem which takes in four main components such as posteriority, prior, likelihood, and evidence. Clustering methods such as K-means, DBSCAN were used. Kmeans cannot handle data with outliers and is prone to fail. Thus K Medoids which is used to find k clusters in n objects by finding the representative of each cluster. It is used to handle data with outliers but is not suitable for high dimensional data. DBSCAN is more efficient and it performs better since the clustering is based on density of distance between objects. Evaluation of data for both the methods were done and it was found that Random Forest had the highest accuracy with 91% and an error rate of 8%. It also showed that class prediction with 91.23% for class no and 85.53% for class yes. For clustering K-medoids outperformed other models with an accuracy of 96% and an error rate of 3% with a k =2 and a 1:23 ratio which is promising. This provides various methods for focusing the marketing campaign on potential customers and making it better (Alexandra & Sinaga, 2021).

This paper titled “Bank Directing Marketing Analysis of Data Mining Techniques’’ deals with the development of successful models to predict the best campaign contact for client subscribing deposit. They mention the usage of Multilayer Perceptron Neural Network (MLPNN), Tree augmented Naive Bayes (TAN), Nominal or Logistic Regression (LR), and Ross Quinlan’s new decision tree model (C5.0). They use the bank direct marketing dataset from the UC Irvine Machine Learning Repository, containing nominal and numerical

attributes. The data is from a Portuguese banking institution's direct marketing campaigns, with the marketing campaigns being based on phone calls. The best campaign turns out to be the one with the highest response rate, which determines how many customers responded positively to the marketing campaign.

There have been varied results by each of the experiments, with each having almost similar testing and training accuracies. The MLPNN model gave an accuracy of 90.79% and 90.02%, the TAN model with an accuracy of 89.16% and 88.75%, the LR model giving 90.09% and 90.43%, while the C5.0 having these values as 93.23% and 90.09%, each for training and testing respectively, showing that C5.0 model achieves effective prediction than the others (A Elsalamony, 2014).

There were other significant patterns observed in the dataset, with 'Duration' being the most significant predictor of all the models. The gains charts for all the models also show that all the models except TAN had similar success, with TAN staying far away from the others. Specificity measures applied on all of the models have also been able to determine that C5.0 was the best in accuracy, sensitivity and specificity analysis for training samples, while similarly for the testing samples, C5.0 had only specificity analysis to be the best from others. MLPNN had the best for accuracy with LR being the best for sensitivity in the same testing samples (A Elsalamony, 2014).

Data Collection and Cleaning:

The dataset is collected as a csv file from data world website with 21 features. It was loaded to google collab using python libraries. In order to proceed for modeling initial data processing and data exploration steps were performed. The sample dataset is loaded as shown below in Figure 1.

Figure 1

Sample dataset

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	campaign	pdays	previous	poutcome	emp_var_rate	cons_price_idx
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.994
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.994
2	37	services	married	high.school	no	yes	no	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.994
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.994
4	56	services	married	high.school	no	no	yes	telephone	may	mon	...	1	999	0	nonexistent	1.1	93.994
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
41183	73	retired	married	professional.course	no	yes	no	cellular	nov	fri	...	1	999	0	nonexistent	-1.1	94.767
41184	46	blue-collar	married	professional.course	no	no	no	cellular	nov	fri	...	1	999	0	nonexistent	-1.1	94.767
41185	56	retired	married	university.degree	no	yes	no	cellular	nov	fri	...	2	999	0	nonexistent	-1.1	94.767
41186	44	technician	married	professional.course	no	no	no	cellular	nov	fri	...	1	999	0	nonexistent	-1.1	94.767
41187	74	retired	married	professional.course	no	yes	no	cellular	nov	fri	...	3	999	1	failure	-1.1	94.767

41188 rows × 21 columns

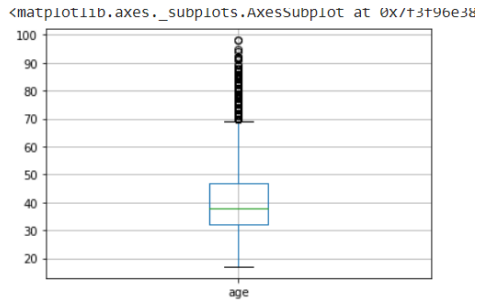
Data Exploration Steps

Check Outliers

For each column outliers were verified by plotting histograms and boxplots. The Figure 2 below shows outliers for the age column where the age above 70 and below 9 are considered as outliers considering most of the customers of that age will not be subscribed for term deposit. Hence replacing those values with the median value.

Figure 2

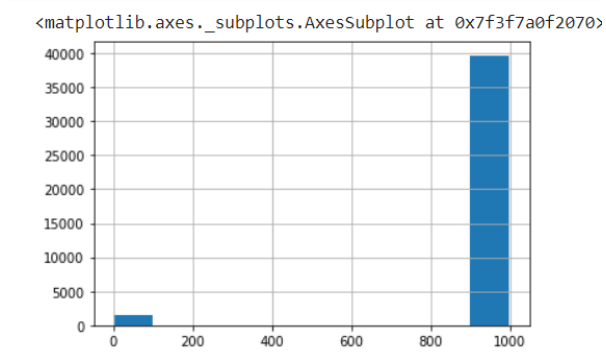
Boxplot of age column



Similarly for the numerical columns such as duration,campaign the outliers were replaced with the median value of the respective columns. By plotting histogram for the column's pdays, it is observed that 96% of the column have similar value 999 which indicates that previous customers are not contacted which can be seen in below Figure 3.

Figure 3

Histogram for pdays column



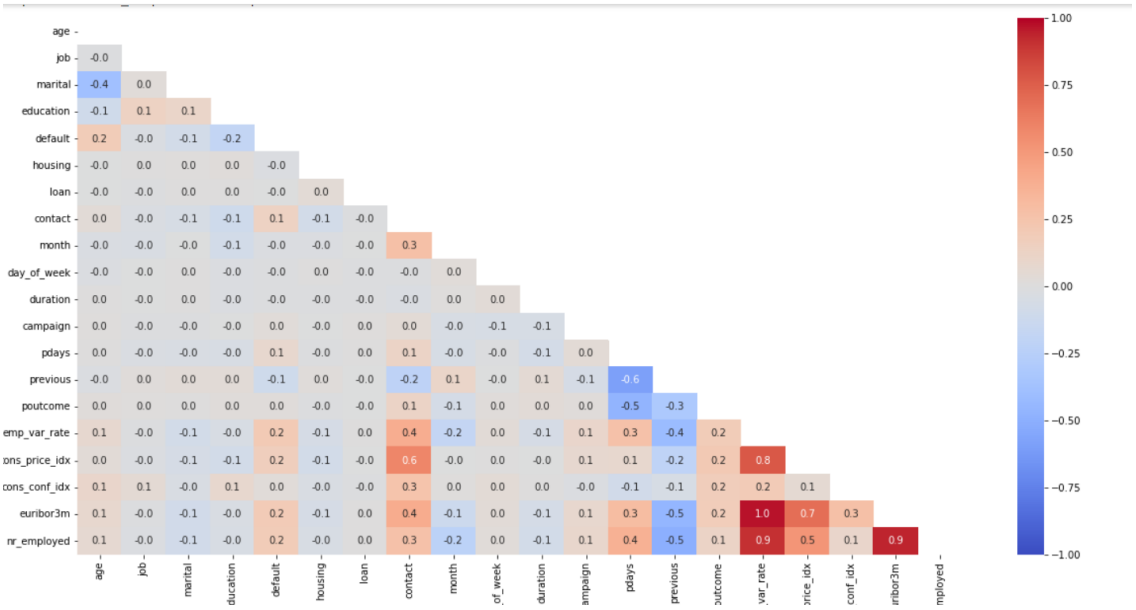
The columns previous, poutcome, emp_var_rate, cons_price_idx, euribor3m, nr_employed, cons_conf_idx have been plotted histograms and boxplots to verify outliers and imbalance data. And it is found that most of the columns have balanced data with minimal outliers which can be used for further modeling.

Transforming categorical to numerical columns

Using Labelencoder, the categorical columns Job, Marital, Education, Default, Housing, Loan, Contact, month, day_of_week were transformed to numerical columns.Once all the columns are transformed a correlation matrix is plotted to find the relation between various features. From the Figure 4 below it is shown that the columns emp_var_rate, cons_price_idx, euribor3m, nr_employed are highly correlated.

Figure 4

Correlation Matrix for Banking Dataset



The final transformed dataset is used for feature selection methods and modeling purposes.

Feature Selection Methods

Feature selection is the process of eliminating unnecessary features and opting out the best features to train an effective machine learning model. It is preferable to limit the number of input variables in order to reduce modeling computational costs and in certain situations, increase model performance. Statistical-based feature selection approaches entail applying statistics to evaluate the relationship between each input variable and the target variable and selecting the input variables having the strongest link with the target variable. Although the choice of statistical measures is dependent on the data type of both the input and output variables, these approaches can be quick and successful. Thus it is challenging to get the appropriate statistical measure for the dataset when we are performing feature selection. The predictive process slows down in the presence of a large number of variables and also requires a lot of memory. There are two main types of feature selection which is supervised and unsupervised feature selection. These filter-based feature selection methods employ statistical measures which compute the score of correlation between the variables present in the input. In this project, we use several feature selection methods such as Logistic Regression and Chi-Square Test.

Chi-Square Test:

Chi-Square Test is employed to examine the independence of the attributes. This method computes the difference between the predicted and the actual response. The variables having less chi-square value are less dependent on each other and are strongly correlated when having higher values. Initially, the attributes are assumed to be independent which forms the null hypothesis. It tests the relationship between the features. It can also be used to test the independence of two events. It is among the most commonly used nonparametric tests.

The formula for the chi-square test is given as

$$X^2 = \Sigma \frac{(O-E)^2}{E}$$

Where X^2 is the chi-square test static

Σ is the summation operator

O is the observed frequency and
E is the expected frequency

When the difference between the observed and the expected value is large the chi-square value is large. To check whether the difference is large we compare the chi-square value to a critical value. The chi2 function from the sklearn.feature selection package was used to apply this strategy to our dataset. This function was given two parameters: all of the descriptive feature data and the array of the target feature. The most significant features, according to the findings of this chi-squared test, are contact, default, education, outcome, job, campaign, age, marital, day_of_week, housing, cons_price_idx, month, and loan. This is done by considering a confidence level of 90% then columns age, job, housing, loan, campaign, and previous are eliminated from the consideration.

Logistic regression method:

In the Logistic Regression(Logit) model, the features are selected based on the p-value score of the feature. Logistic regression is the type of regression analysis that is used to find the probability of a certain event which is occurring. It is best suited for cases where we have a categorical dependent variable that can take only discrete values. The binary (dichotomous) response variable (e.g., 0 and 1, true and false) is modeled as a linear combination of a single or more independent variables in logistic regression.

Univariate logistic regression involves just one independent variable, whereas multivariate logistic regression involves many independent variables.In logistic regression, the response variable's probability or chances (rather than its values, as in linear regression) are represented as a function of the independent variables. Variables such as odds has the probability of whether an event will occur or not. It is given by (p/1-p) and it can range from 0 to infinity. The odds ratio (OR) is the difference between two odds. OR values can vary from 0 to positive infinity. OR is important in evaluating regression coefficients, i.e. the influence of independent variables on the response variable, because regression coefficients are difficult to read. OR may be calculated by exponentiating regression coefficients. It is critical to include the best independent variables (features) for fitting the regression model in order to make accurate predictions of the regression outcome (e.g. variables that are not highly correlated). If you include all characteristics, you might not receive all significant predictors in the model. The features that are having a p-value less than 0.05 are assumed to be more relevant.

In our dataset we are considering a 90% confidence level then the columns age, job, housing, loan, campaign, and previous are eliminated from the consideration. Other than the previously mentioned columns, all columns are considered and make input data including those columns.

Machine learning models used

K -Nearest Neighbour

The k-nearest neighbors (KNN) algorithm is a data categorization approach that estimates the chance that a data point will belong to one of two groups depending on which data points are closest to it. It is a supervised machine-learning algorithm that can perform both classification and regression. KNN is also used for missing value imputation. KNN performs the classification assuming that the nearest data points to the given data point are similar and considered to be of the same class. A higher k value results in smoother separation curves, resulting in less complex models. Smaller k values, on the other hand, tend to overfit the data,

resulting in complex models.

When analyzing a dataset, it is important to select the correct k-value in order to avoid overfitting and underfitting. It is known as a lazy learning algorithm or lazy learner since it does not do any training when given training data. Instead, it just saves the data throughout the training period and makes no computations. It does not create a model until a query is run on the dataset. As a result, KNN is great for data mining. Simply said, KNN attempts to establish the group to which a data point belongs by examining the data points around it. One of the most notable advantages of employing the KNN method is that no model or parameters must be built or tuned. When you have a large volume of data, it might be difficult to extract quick and plain information from it. The k-nearest neighbor's algorithm is highly susceptible to overfitting due to the curse of dimensionality. But with the help of feature selection methods such as Logistic Regression and the Chi-square test we are able to reduce the number of features and thus fit it into the KNN model and the model has quite a bit of improvement in performance. This is shown in the figure below.

KNN Reports

	precision	recall	f1-score	support
False	0.97	0.91	0.94	7310
True	0.91	0.97	0.94	7310
accuracy			0.94	14620
macro avg	0.94	0.94	0.94	14620
weighted avg	0.94	0.94	0.94	14620

The above figure shows the performance of KNN without feature selection. It shows that it has an accuracy of 94% and a good F1 score.

KNN Reports

	precision	recall	f1-score	support
False	0.90	0.92	0.91	7310
True	0.92	0.89	0.91	7310
accuracy			0.91	14620
macro avg	0.91	0.91	0.91	14620
weighted avg	0.91	0.91	0.91	14620

The above figure shows the performance of KNN after doing feature selection using logistic regression and it is found that it has a good precision and recall score of around 92%. The accuracy is around 91%.

KNN Reports

	precision	recall	f1-score	support
False	0.78	0.89	0.84	7310
True	0.88	0.75	0.81	7310
accuracy			0.82	14620
macro avg	0.83	0.82	0.82	14620
weighted avg	0.83	0.82	0.82	14620

The above figure shows the performance of KNN after the selection of features using the chi-square test and it is found that the accuracy has dropped from 91% to 82%. This might be due to the fact that due to the sample size requirements and difficulty to interpret when there are a large number of categories there occurs a tendency that provides low correlation measures even for highly significant results.

SVM

Support Vector Machine (SVM) is widely used in data analysis and pattern recognition for classification and regression analysis. SVM can perform both linear and non-linear classification. The non-linear classification can be performed using various kernel functions like sigmoid, hyperbolic tangent function, gaussian, etc. The cross-Validation method can be used in selecting the appropriate parameters.The support vector machine algorithm seeks a hyperplane in N-dimensional space (N — the number of features) that distinguishes between data points.

There are several hyperplanes that might be used to split the two groups of data points. Maximizing the margin distance gives some reinforcement, allowing subsequent data points to be categorized with more certainty. Hyperplanes are decision boundaries that aid in the classification of data items. Data points on either side of the hyperplane might belong to distinct classes. Furthermore, the size of the hyperplane is determined by the number of features. When the number of input features is two, the hyperplane is just a line. When the number of characteristics exceeds a certain number, it becomes impossible to imagine.

So in order to get a perfect hyperplane only the important features are to be selected.It can be done with a variety of feature selection techniques such as the ones we have considered in our project which are logistic regression and chi-squared test.

SVM Reports

	precision	recall	f1-score	support
False	0.51	0.50	0.50	7310
True	0.51	0.52	0.51	7310
accuracy			0.51	14620
macro avg	0.51	0.51	0.51	14620
weighted avg	0.51	0.51	0.51	14620

The above figure shows that without any feature selection methods it performs with an accuracy of 51%.

SVM Reports

	precision	recall	f1-score	support
False	0.50	0.53	0.52	7310
True	0.50	0.48	0.49	7310
accuracy			0.50	14620
macro avg	0.50	0.50	0.50	14620
weighted avg	0.50	0.50	0.50	14620

The above figure shows the precision, recall, and accuracy of SVM after the selection of features using chi-squared metrics. The SVM has an accuracy of 50% which is the same as the base model but the number of features is less which helps in memory efficiency.

SVM Reports

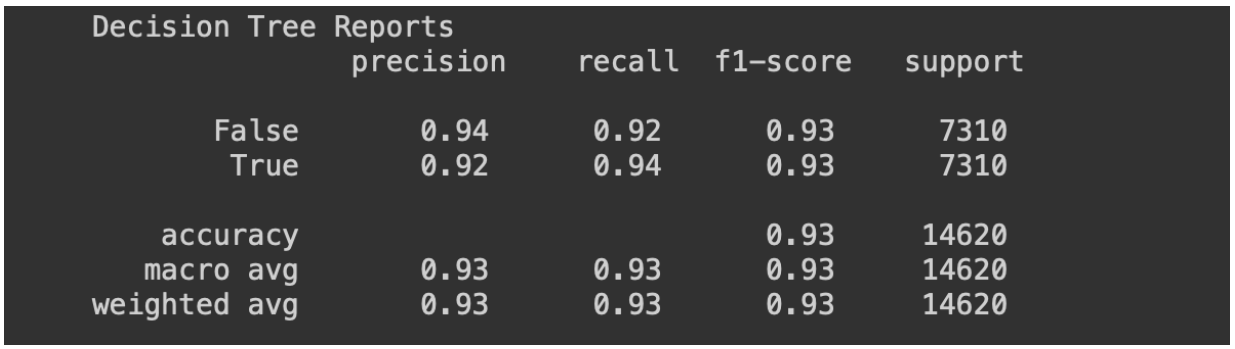
	precision	recall	f1-score	support
False	0.59	0.43	0.50	7310
True	0.55	0.71	0.62	7310
accuracy			0.57	14620
macro avg	0.57	0.57	0.56	14620
weighted avg	0.57	0.57	0.56	14620

The figure above shows the precision, recall, and accuracy of SVM after the selection of features using chi-squared metrics. The SVM has an accuracy of 57% which is significantly higher than other methods and the baseline model also has a lesser number of features.

Decision Tree Classifier

Decision Tree Classifier is a supervised machine learning model which can perform both classification and regression. It is simple and easy to understand by looking at the tree. This method consists of 3 types of nodes: Root Node, Internal Node and Leaf Node. Each node in a decision tree is assigned an attribute based on its ability to reduce chaos in the classification. Decision trees can perform both classification and regression tasks, CART algorithm: Classification and Regression Tree. The idea behind Decision Trees is that you utilize the dataset characteristics to build yes/no questions and then partition the dataset until you isolate all data points from each class. Every time you ask a question, you add a node to the tree. And the first node is known as the root node. The last nodes are called the leaf nodes.

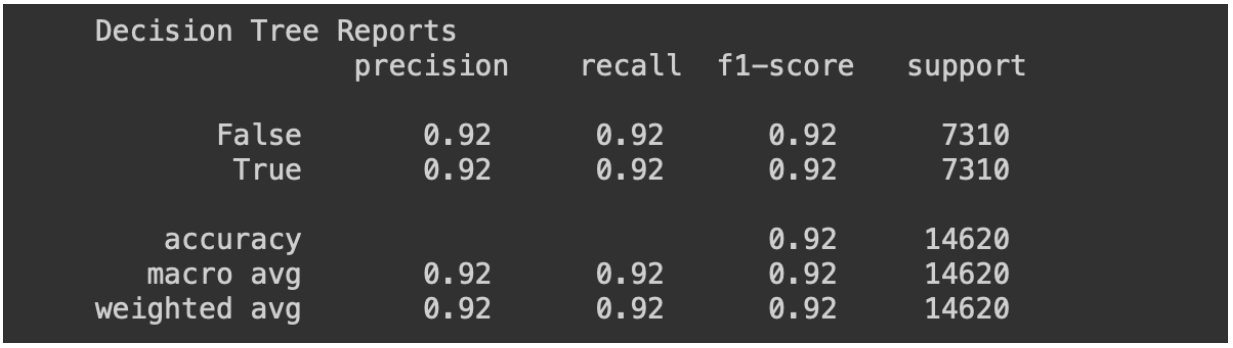
Instead of attempting to make the best overall decision, a greedy method makes locally optimum judgments to select the feature utilized in each split. Because it optimizes for local decisions, it is solely concerned with the node at hand and what is optimal for that node in particular. As a result, it is not necessary to investigate all possible splits for that node and beyond. Information gain is calculated from gini and entropy.



Decision Tree Reports

	precision	recall	f1-score	support
False	0.94	0.92	0.93	7310
True	0.92	0.94	0.93	7310
accuracy			0.93	14620
macro avg	0.93	0.93	0.93	14620
weighted avg	0.93	0.93	0.93	14620

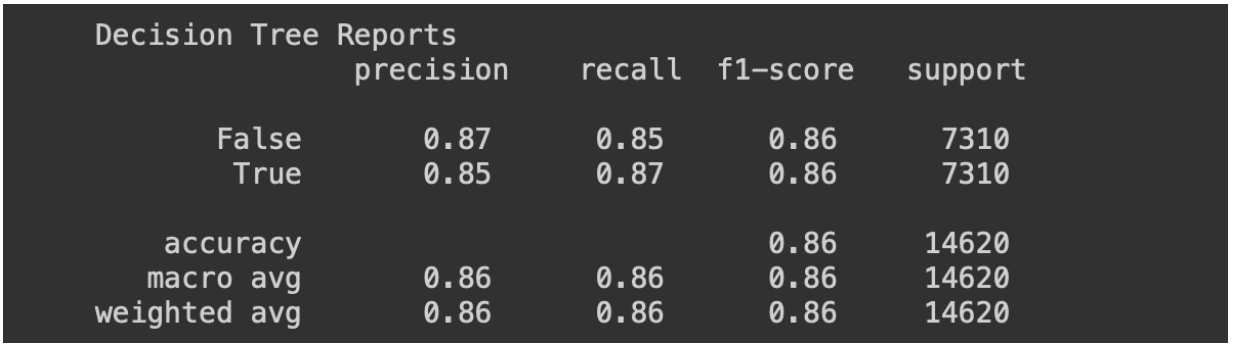
The above figure shows the decision tree classifier without any feature selection methods performs well with an accuracy of 93% due to data robustness since the algorithm handles all types of data in a good manner.



Decision Tree Reports

	precision	recall	f1-score	support
False	0.92	0.92	0.92	7310
True	0.92	0.92	0.92	7310
accuracy			0.92	14620
macro avg	0.92	0.92	0.92	14620
weighted avg	0.92	0.92	0.92	14620

The above figure shows how the Decision tree performs after feature selection by logistic regression. The accuracy is still the same but this was done with a lesser number of features.



Decision Tree Reports

	precision	recall	f1-score	support
False	0.87	0.85	0.86	7310
True	0.85	0.87	0.86	7310
accuracy			0.86	14620
macro avg	0.86	0.86	0.86	14620
weighted avg	0.86	0.86	0.86	14620

The above figure shows the performance of the decision tree after feature selection using the chi-squared test. It reduced the accuracy a bit from 93% to 86%.

Random Forest

Random forest is a supervised machine learning model which can perform both classification and regression tasks. A random forest produces accurate, understandable predictions. It is capable of efficiently handling large datasets. In terms of prediction accuracy, the random forest method performs better than the decision tree approach. Random forest performs well on the combination of numerical and categorical data. It requires a lot of computational power as well as resources because it builds a lot of trees and then combines their outputs. It consists of a large number of individual decision trees that can operate as an ensemble. The key is the low correlation between models. Uncorrelated models have the ability to provide ensemble forecasts that are more accurate than any of the individual predictions, just like assets with low correlations (like stocks and bonds) combine to form a

portfolio that is greater than the sum of its parts. The trees shield each other from their own mistakes, which results in this amazing effect. It follows 2 main methods such as Bagging and Feature randomness.

Bagging (Bootstrap Aggregation) : Because of how sensitive trees are to the data they are trained on, even minor adjustments to the training set can produce dramatically different tree architectures. By enabling each individual tree to randomly sample from the dataset with replacement and produce various trees as a consequence, random forest takes advantage of this. This procedure is referred to as bagging.

Feature randomness : When splitting a node in a typical decision tree, we analyze all potential features and choose the one that results in the greatest gap between the observations in the left node and those in the right node. In the end, this leads to less correlation between trees and increased diversity by forcing even more variety across the model's trees.

Random Forest Reports

	precision	recall	f1-score	support
False	0.96	0.94	0.95	7310
True	0.94	0.97	0.95	7310
accuracy			0.95	14620
macro avg	0.95	0.95	0.95	14620
weighted avg	0.95	0.95	0.95	14620

The above figure represents the random forest methods and shows that it gives an accuracy of 95%.

Random Forest Reports

	precision	recall	f1-score	support
False	0.95	0.93	0.94	7310
True	0.93	0.95	0.94	7310
accuracy			0.94	14620
macro avg	0.94	0.94	0.94	14620
weighted avg	0.94	0.94	0.94	14620

The above figure represents the random forest methods with feature selection (Logistic regression) and shows that it gives an accuracy of 94% which is similar when compared to the baseline model but with less number of features.

Random Forest Reports

	precision	recall	f1-score	support
False	0.90	0.90	0.90	7310
True	0.90	0.90	0.90	7310
accuracy			0.90	14620
macro avg	0.90	0.90	0.90	14620
weighted avg	0.90	0.90	0.90	14620

The above figure represents the random forest model with chi-square feature selection methods. Since it may be seen that some of our important features which gives some

information to our model are eliminated in chi-square method the accuracy drops to 90%.

Logistic Regression

Logistic Regression is a supervised machine learning algorithm that is used to predict an event occurrence. It is a statistical method for describing and explaining the relationship between one dependent binary variable and one or more independent variables. This model helps in identifying the various anomalies in the data which are predictive of fraud. Logistic regression is having its applications in various fields like manufacturing, marketing, finance, healthcare industry, etc.

There are three types of logistic regression which are defined as follows based on the categorical responses such as Binary logistic regression, multinomial, ordinal logistic regression.

Binary logistic regression: In this approach, the response or dependent variable is dichotomous in nature—i.e. it has only two possible outcomes (e.g. 0 or 1). This method is the most widely applied in logistic regression, and it is also one of the most widely used classifiers for binary classification in general.

In our project we will be applying binary logistic regression and the results are as follows.

Logistic Regression Reports

	precision	recall	f1-score	support
False	0.77	0.82	0.80	7310
True	0.81	0.75	0.78	7310
accuracy			0.79	14620
macro avg	0.79	0.79	0.79	14620
weighted avg	0.79	0.79	0.79	14620

The above figure shows the Logistic regression. It can be seen that it gave us an accuracy of 79% with a low precision score compared to other methods.

Logistic Regression Reports

	precision	recall	f1-score	support
False	0.75	0.81	0.78	7310
True	0.79	0.72	0.76	7310
accuracy			0.77	14620
macro avg	0.77	0.77	0.77	14620
weighted avg	0.77	0.77	0.77	14620

The above figure shows the Logistic regression after applying logistic regression methods for feature selection. It can be seen that it gave us an accuracy of 77% with a low precision score compared to other methods. But only the important features were selected.

Logistic Regression Reports

	precision	recall	f1-score	support
False	0.75	0.65	0.70	7310
True	0.69	0.79	0.74	7310
accuracy			0.72	14620
macro avg	0.72	0.72	0.72	14620
weighted avg	0.72	0.72	0.72	14620

The above figure shows the Logistic regression with the chi-square feature selection method. The accuracy has dropped considerably when compared to the logistic regression method and also the total number of features selected is also considerably less.

XGBoost

XGBoost stands for extreme gradient boosting which is an implementation of gradient-boosting trees. In gradient boosting, every predictor corrects its predecessor's errors. In XGBoost, decision trees are created in a sequential form where the weights play an important role. All of the independent variables are given weights, which are then fed into the decision tree to predict the outcome. The incorrectly predicted weight of variables are then increased and fed into the second decision tree. These individual classifiers are then combined to form an effective and precise model.However, decision tree-based algorithms are currently thought to be best-in-class for small- to medium-sized structured/tabular data.It can be used for many different applications: can be used to resolve challenges involving regression, classification, ranking, and custom prediction. Both XGBoost and Gradient Boosting Machines (GBMs), ensemble tree approaches, use the gradient descent architecture to boost weak learners (CARTs in general). But XGBoost enhances the fundamental GBM architecture with system optimization and algorithmic improvements. It uses three main techniques such as parallelization, tree pruning and hardware optimization.

In our cases when we applied XGBOOST to our dataset we obtained the following results.

XGBoost Reports

	precision	recall	f1-score	support
False	0.94	0.89	0.91	7310
True	0.89	0.95	0.92	7310
accuracy			0.92	14620
macro avg	0.92	0.92	0.92	14620
weighted avg	0.92	0.92	0.92	14620

The above figure shows the value of XGBOOST without any feature selection. When we consider the accuracy we can observe that the accuracy is around 92%.

XGBoost Reports

	precision	recall	f1-score	support
False	0.93	0.87	0.90	7310
True	0.88	0.94	0.91	7310
accuracy			0.90	14620
macro avg	0.91	0.90	0.90	14620
weighted avg	0.91	0.90	0.90	14620

The above figure shows the value of XGBOOST with logistic regression feature selection. Even Though it is not a big change in accuracy it does it with a lesser number of features which is good for memory efficiency.

XGBoost Reports

	precision	recall	f1-score	support
False	0.80	0.87	0.84	7310
True	0.86	0.79	0.82	7310
accuracy			0.83	14620
macro avg	0.83	0.83	0.83	14620
weighted avg	0.83	0.83	0.83	14620

The above figure shows the value of XGBOOST with chi-square test feature selection. It can be seen that there is a significant drop in accuracy and also the precision of True drops to a lower point.

Results

We trained various machine learning models like KNN, SVM, Decision Tree Classifier, Random Forest, Logistic Regression, and XGBoost and compared them using the evaluation metrics like accuracy score. The models were trained using various feature selection methods like the Logistic Regression method and Chi-Square Test method. Also, we performed the model training without the feature selection methods and compared them with the models trained using the feature selection methods and compared them using different evaluation metrics like accuracy, precision, sensitivity/recall, and f1-score.

Accuracy gives the information of how many times the machine learning model gave the correct results in prediction. It is the ratio of the number of correct predictions in the total predictions. The accuracy can be solved using the various values obtained from the confusion matrix which includes True Positives(TP), True Negatives(TN), False Positives(FP) and False Negatives(FN). True Positives is the number of customers who took term deposits and the model also predicted customers who took the term deposit whereas True Negative is the number of customers who did not take term deposits and the model also predicted the same. False Positives are the customers who did not take the term deposit and the model predicted the customer took the term deposit whereas False Negatives are the customers who took the term deposit but the model predicted the customer did not take the term deposit. Accuracy can be calculated by the following formula

Accuracy = Correct Predictions/a Total number of Predictions

Accuracy = $(TP+TN)/(TP+TN+FP+FN)$

Precision is a measure of the number of the positive classes predicted that actually belong to the total positive classes. It is the ratio of the truly positive among all the positives. Precision can be calculated by the following formula

Precision = Truly Positive / Total Positives

Precision = $TP/(TP+FP)$

Recall is the measure of positive class predictions in the total positives in the dataset. It is the number of correct positive predictions made out of all correct positives in the dataset. If the recall is high, the model can predict the positive samples as positive. Recall can be calculated by the following formula

Recall = $TP/(TP+FN)$

F1-Score is the harmonic mean of precision and recall. It can be calculated by the following formula
F1-Score=(2*Precision*Recall)/Precision+Recall

The table below shows the accuracy scores obtained using the Chi-Square Test vs no feature selection method vs Logistic Regression.

The results observed for various models, with or without feature selection, are compared. We perform these models without the feature selection, and then with it, once using the technique “Chi-square Test(CT)” and the other time using “Logistic Regression(LR)”. Each model’s statistics differ with each feature selection method, and we observe the following.

Table 3

Accuracy Scores of Chi-Square Test, Logistic Regression, and No-Feature Selection Methods.

Accuracy Score	KNN	SVM	Decision Tree Classifier	Random Forest	Logistic Regression	XGBoost
Without Feature Selection	94%	51%	93%	95%	79%	92%
Using Chi-Squared Test	82%	57%	86%	90%	72%	83%
Using Logistic Regression	91%	50%	92%	94%	77%	90%

The accuracy is tabulated above for all the measures in the table. The highest accuracy was found for the Random Forest model without utilizing any feature selection for the data, at about 95%. The least found was with SVM but where the feature selection was made by Logistic Regression. For all the models except SVM, the accuracy was above 70%, and for models except Logistic Regression and SVM, there was an accuracy above 80% observed. The accuracy is observed to be more in the models without any feature selection methods as we consider all the features of the dataset. When we are training the models with the feature selection methods, it was observed that all the models except SVM were having more accuracy in the Logistic Regression feature selection method when compared to the Chi-Squared test method. Models such as Decision tree classifier, Random forest, XGBoost have comparable accuracy with feature selection using logistic regression and it is memory efficient because there is less number of features

Table 4

Precision values of various models with and without using feature selection

Precision	KNN	SVM	Decision Tree Classifier	Random Forest	Logistic Regression	XGBoost
Without Feature Selection	91%	51%	92%	94%	81%	94%
Using Chi-Squared Test	88%	55%	85%	90%	69%	86%
Using Logistic Regression	92%	50%	92%	93%	79%	88%

The precision of the models is observed in the table given. Highest precision is seen with both XGBoost and Random Forest models without the use of feature selection on the data, at a value of 94%. As observed with accuracy, all the models except SVM tend to be consistent, with most of them having a recall rate over 85%. Even with the usage of both the feature selection methods, the models tend to perform similarly, with the features selected by Logistic Regression performing better than the features selected by Chi-squared test, except for SVM, where the latter performs better. The precision for the Random Forest model is the highest for all the features, scoring 90% and above for individual features selected. KNN and XGBoost follow thereby, with KNN having the least for Chi-square selected features at 88% and XGBoost similarly at 86%. Only the SVM model has restricted itself to the precision values under 50%-55%. Other models, like the KNN, the Decision Tree and the XGBoost models, have a similar precision rate, for all kinds of features chosen.

Table 5

Recall values of various models with and without using feature selection

Recall	KNN	SVM	Decision Tree Classifier	Random Forest	Logistic Regression	XGBoost
Without Feature Selection	97%	52%	94%	97%	75%	95%
Using Chi-Squared Test	75%	71%	87%	90%	79%	79%

Using Logistic Regression	89%	48%	92%	95%	72%	94%

The recall rate can be seen as the observed table given. Chi-squared Test seems to choose features which have underperformed on recall, except with the Logistic Regression model, where it performs better than others at 79%. The highest observed recall rate can be seen with the KNN and Random Forest models at 97% without any selected features. This is followed by XGBoost at 95%, also seen with the Random Forest classifier where the features are selected by Logistic Regression. The least observed is for the SVM model, where the features are chosen by Logistic Regression, at 48%. Among the feature selected data, the models perform better with Logistic Regression as the feature selection method. The features selected by the Logistic Regression perform almost similar to the original features in all the models, which shows that the recall rate is preserved with the features. The Decision Tree and Random Forest models perform consistently among all other models with or without selected features, with the Decision Tree achieving a minimum of 87% and the Random Forest models achieving over 90%.

Table 6

F1-Score values of various models with and without using feature selection

F1-Score	KNN	SVM	Decision Tree Classifier	Random Forest	Logistic Regression	XGBoost
Without Feature Selection	94%	51%	93%	95%	78%	92%
Using Chi-Squared Test	81%	62%	86%	90%	74%	82%
Using Logistic Regression	91%	49%	92%	94%	76%	91%

The F1-scores are calculated and presented as given. We can observe that the highest value is obtained with the Random Forest model at 95%. This, like with all the other measures, has been consistent with having the best results without the selection of any features. Consequently, the F1-scores of the models without any features selected, have been higher than those of any features selected, except for the SVM model. The least F1-score can be observed with the SVM model when it uses Logistic Regression for the feature selection. Out of all the models, Logistic Regression has identical values for all the features, indicating that the F1-score remained consistent across all features. Random Forest performs well too, with an F1-score at 90% and above for all the features chosen. KNN, Decision Tree Classifier and XGBoost models all perform almost identically, with similar values for the selected features across all of them. This can also show us that the features selected by Logistic Regression are

successful to produce an F1-score similar to the models trained by all the selected features.

Significance of methods :

The two main methods we implemented for our dataset are the Logistic regression feature selection method and chi-square feature selection method. If we compare the results of accuracy and precision in our case, the logistic regression feature selection method is working comparatively better than the other feature selection methods. We are considering accuracy and precision in our case because precision gives us the values of true positives in our project. We need to find the number of people who will opt for long term deposits with the bank. It is best suited for cases where we have a categorical dependent variable that can take only discrete values. The binary (dichotomous) response variable (e.g., 0 and 1, true and false) is modeled as a linear combination of a single or more independent variables in logistic regression. This method works well with our data because it is a parametric approach which has regression coefficients which penalizes more features with non zero coefficients.

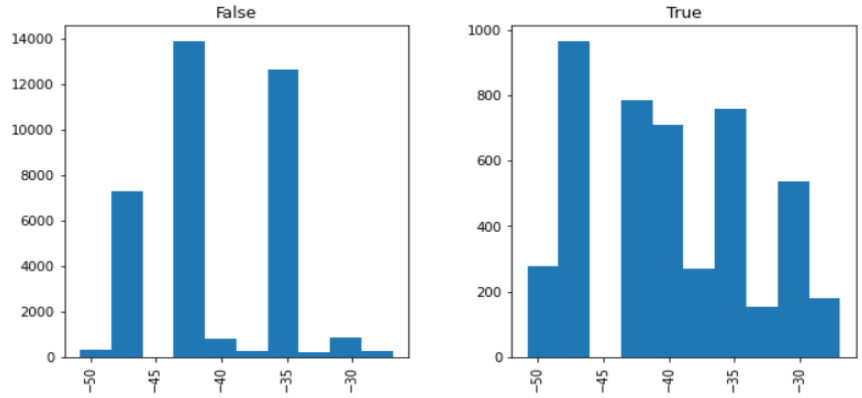
Chi- Square method on the other hand is a useful statistical test to look at the differences between categorical variables. It has a limitation factor which limits the data as well as it becomes difficult to interpret the data when there are large categories of data around 15 features. It is relatively a simple method when compared with logistic regression and should not be used unless the data are independent. In our cases since the data is dependent on each other, features such as education and job are dependent on each other, job and income are dependent on each other etc. When there are various cells with near zero expectation the chi square test fails. In our case even though our p value was good when calculating the equally weighted mean square test we got a very less p value of around 0.040. So the zeros in the data are acting as noise in our case which gives low chi -square value. Advantages of logistic regression on our dataset so that it is less inclined to overfitting in high dimensional datasets and also can effectively interpret features models coefficients for feature importance. It is also very fast at classifying data. It works well with dependent data and since our dataset some of the features are dependent on each other it can interpret them effectively. It also works well with the Random forest algorithm since it is an ensemble method which uses averaging to improve the predictive accuracy and control over-fitting.

Methods such as KNN performed low on our dataset since we know that KNN works well with a smaller number of input variables but struggles when it has a large number of input variables. In our case the number of input variables is high since we have more features. Random forest is an ensemble learning method so is performing better than all the other models since it can do numerous splits at a single node, resulting in greater precision and accuracy. It also determines the statistical significance between sub-nodes and the parent node. It is building multiple trees and ensemble them to get the best performance while being robust to unbalanced data. Random forest with logistic regression is performing comparatively better with the baseline model that too with lesser number of features.The second best performing model was decision tree with logistic regression. The decision tree method gives high importance to a particular set of features in the dataset. It is faster to train but it fails to generalize the data better. XGBoost on the other hand performed good since it takes into account the weight calculations and the feature selection is done internally and hence we don't see much difference in the accuracy.

Difference between all the features and selected features

Feature Selection Method 1 (Chi-Square):

Using this method we filtered the columns in the dataset from 20 (excluding the target column) to 13 columns. We considered the columns with a p-value greater than 0 to filter them. Below are the columns we filtered from the original dataset:
'Cons_conf_idx', 'emp_var_rate', 'nr_employed', 'previous', 'days', 'euribor3m', and 'duration'



‘cons_cong_idx’ is an indicator of the household development of a particular country/city. In this data, there is no information about the country/city considered. That information will be helpful to evaluate the importance of this column. If you see the distribution curves of the ‘cons_cong_idx’ column with the target column wise then there is not much difference in the distribution which can be the reason for not affecting the performance of the model with absence of this column. Similarly, other columns ‘days’, ‘duration’ also are with similar distributions for two target values which is not affecting much with the model performance.

'emp_var_rate' is rate of employment calculated/updated quarterly, ‘nr_employed’ is the number indicating the number of employees in the market. These are very general criteria to analyze the overall trend of the term deposit in the market. But we can’t use these columns as the criteria for predicting the chance of a term deposit of a specific customer. So, this is the reason for showing least importance by this feature selection method.

This method selected ‘contact’, ‘default’, ‘education’, ‘poutcome’, ‘job’, ‘campaign’, ‘age’ and remaining columns in the provided importance. This is the expected column according to the domain importance. Because, when we are considering a customer then ‘default’ tells us the credibility of the customer which is very important, ‘poutcome’, ‘education’, ‘job’ is important which indicates the capability of a customer to bare the term deposit. These columns are inter-related to each other. ‘Age’ is also important (domain knowledge) which indicates the income of the customer indirectly.

Feature Selection Method 2 (Logistic Regression):

By using this method, we applied a 90% confidence interval and filtered the columns. This method eliminated ‘age’, ‘job’, ‘housing’, ‘loan’, ‘campaign’, ‘previous’ from the original dataset. ‘Campaign’, ‘previous’ are filtered and the reasons for those are already in the previous method. So, it can be the same with this method also. But strangely this method

removed ‘age’, ‘job’, ‘housing’, ‘loan’ which we thought were important features according to the domain knowledge. But later we realized that we tested the columns with the original dataset which is not balanced. So, there is not much data for true values in the dataset to estimate the importance of those columns with respective the target column. With the limited true values data, the model couldn’t find the importance of these columns with the target value. That’s why if we have a balanced dataset, then maybe this could find the importance of these columns.

Interpretation of results (Conclusion)

In this project, Random Forest outperformed (95%) all other models with or without feature selection. This is obvious considering the power of random forest than other normal models. But, this model takes more computational time than others. Random forest with logistic regression feature selection method got accuracy of 94% which is very less depreciate (1%) even with eliminating six columns from the original dataset. Surprisingly, even eliminating the columns which we felt are important according to the domain knowledge, but model performed well.

The next best performed model was Decision Tree Classifier with 93% accuracy (without feature selection) and 92% with feature selection (logistic regression). Accuracy affected with only 1% change even with eliminating six columns. In our datasets, most of the columns are in the form of categories which may be the reason for better performance apart from normal advantages of model

The overall performance of all models with Chi-square feature selection got affected (depreciated) much more than with the other feature selection model. The reason for this as mentioned earlier, is the indirect relationship between the columns in the dataset. This can be the reason for that.

In our project, we used the SVM method as one of the models. Size of the dataset is the major factor for performance of the model. In our dataset, we have 41K rows in the dataset with 20 columns which is a huge size. That’s why SVM is performing poorly among all the models. But somehow there is improvement in the performance with feature selection (chi-square).

‘marital’, ‘default’, ‘contact’ are selected in both feature selection methods. This reveals the importance of the columns. So, we can use these columns with priority for some business purposes. We can use this information while we are targeting the customers. suppose , majority of the customers are married who took the term deposit. Therefore we can change the strategy such that to target married customers to increase the term deposits and also need to conduct the survey or any other investigation to analyze the other categories in the marital column. ‘default’ is an important column that indicates the financial status of the customer. So, we need to filter the customers when we are targeting term deposits. According to the results, contact column is also making a difference in the term deposits. Apart from these, columns ‘job’, ‘education’, ‘age’ and etc are making a difference in the decision of term deposit. Therefore will consider these columns to make new strategies or policies to increase term deposits and also can make new offers to attract those sections of customers.

References

A Elsalamony, H. (2014). Bank Direct Marketing Analysis of Data Mining Techniques. *International Journal of Computer Applications*, *85*(7), 12-22.

Bank marketing - dataset by XPRIZEAI-Ai. data.world. (2017, July 19). Retrieved October 21, 2022, from https://data.world/xprizeai-ai/bank-marketing

Çığşar, B., & Ünal, D. (2019). Comparison of data mining classification algorithms determining the default risk. *Scientific Programming*, *2019*, 1–8. https://doi.org/10.1155/2019/8706505

J. Alexandra and K. P. Sinaga, "Machine Learning Approaches for Marketing Campaign in Portuguese Banks," 2021 3rd International Conference on Cybernetics and Intelligent System (ICORIS), 2021, pp. 1-6, doi: 10.1109/ICORIS52787.2021.9649623.

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22-31.

Wang, D. (2020, October). Research on Bank Marketing Behavior Based on Machine Learning. *In Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacture* (pp. 150-154). https://doi.org/10.1145/3421766.3421800