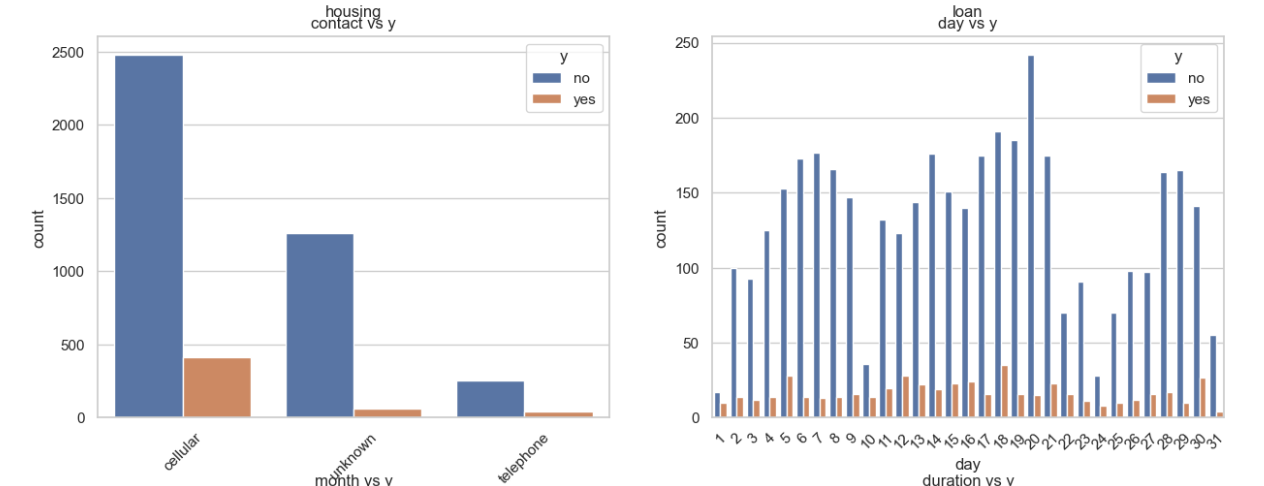
**CREDIT SCORING APPROACHES**

In this report, we will discuss and compare two distinct approaches for credit scoring problem. The first approach involves a Logistic Regression model with feature elimination, while the second approach employs a Bayesian Network model. We will provide an overview of each approach, delve into their key components, and conclude with a comparative analysis of their strengths and limitations. The dataset that is used for both of these models was selected from Kaggle “https://www.kaggle.com/datasets/kapturovalexander/bank-credit-scoring “.

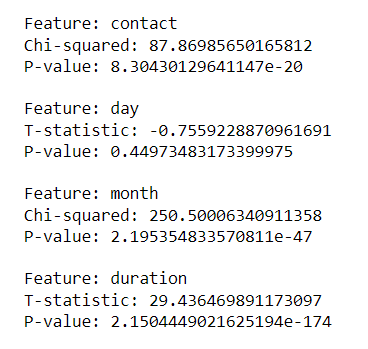
**Approach 1:** Logistic Regression with Feature Elimination

The first approach utilizes a Logistic Regression model to predict credit-scoring outcomes. This method involves the following steps:

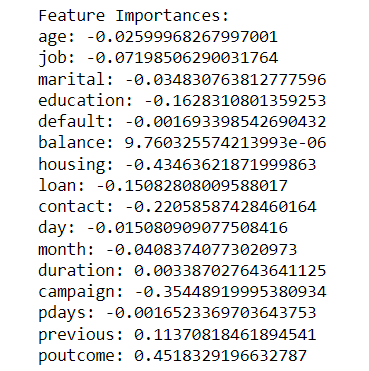
1. Data Preparation and Preprocessing: This step contain initial dataset exploration to identify missing values and understand feature distribution. Then conversion of categorical features to numerical values was done with manual mapping. Lastly, in this section we can include visualization of feature distributions with respect to target variable to gain insights.



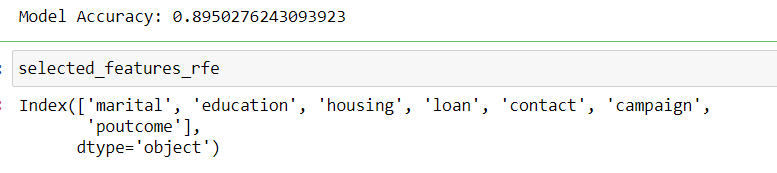
1. Statistical Tests: Here with the help of statistical tests we assess the relationship between each feature and the target variable 'y'. It is essentially checking whether there is a significant difference in the distribution of a feature between the positive ('yes') and negative ('no') classes of the target variable. For categorical features, the code calculates the chi-squared statistic and p-value. A low p-value indicates that the feature is likely associated with the target variable. A high chi-squared value suggests a strong association. As for numeric features, the code performs an independent two-sample t-test between the feature values for the 'yes' and 'no' classes. A low p-value indicates that there is a significant difference in means between the two classes.



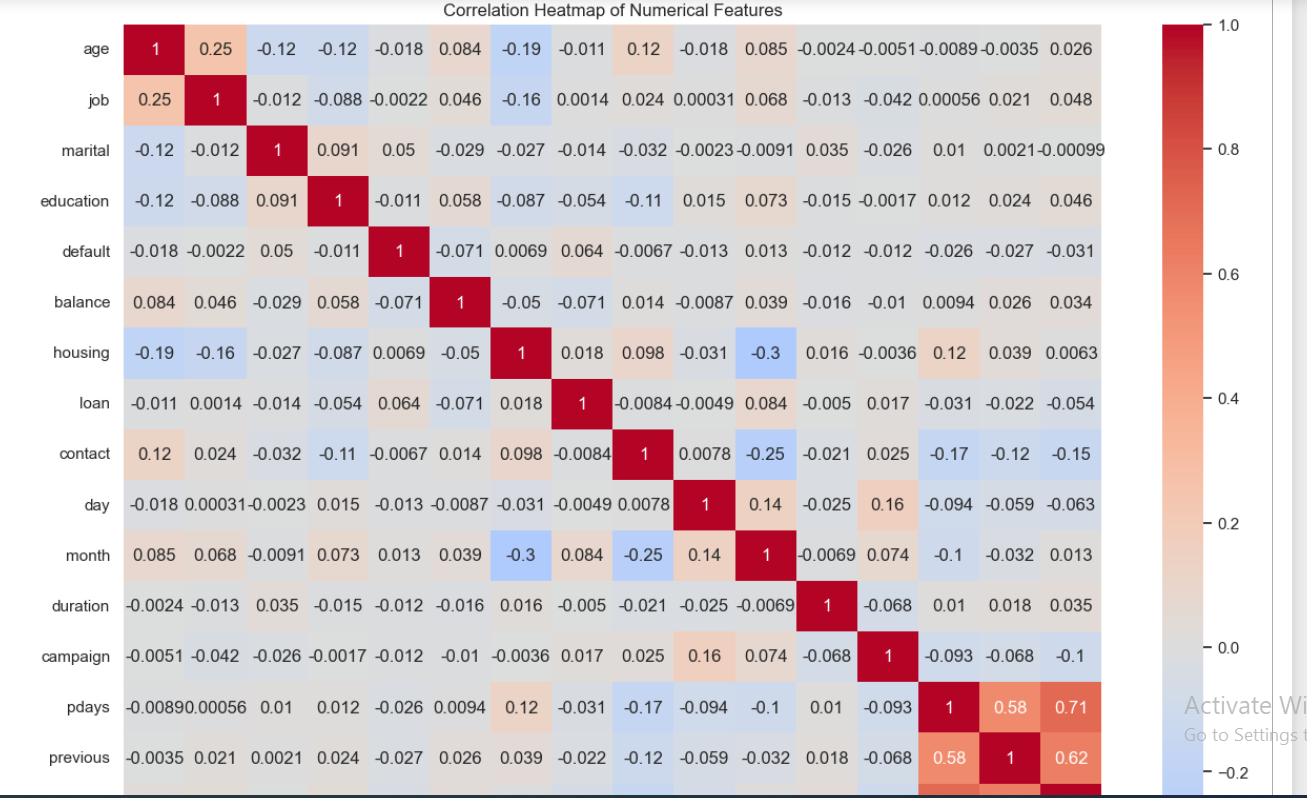
1. Feature Importance Assessment: Next step were feature assessment and here we trained our model on entire dataset and got the coefficients which is also referred as feature importance. Then we analyzed this numbers and identified which feature have more influence in model training.



1. Feature Selection with Recursive Feature Elimination (RFE): For this step, we used RFE to select the most relevant features for the model. Then trained logistic regression model on this selected features. Finally evaluated accuracy with test dataset.



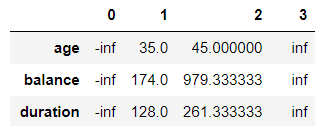
1. Correlation Analysis: In this last step, we did computation and visualization of the correlation matrix among numerical features.



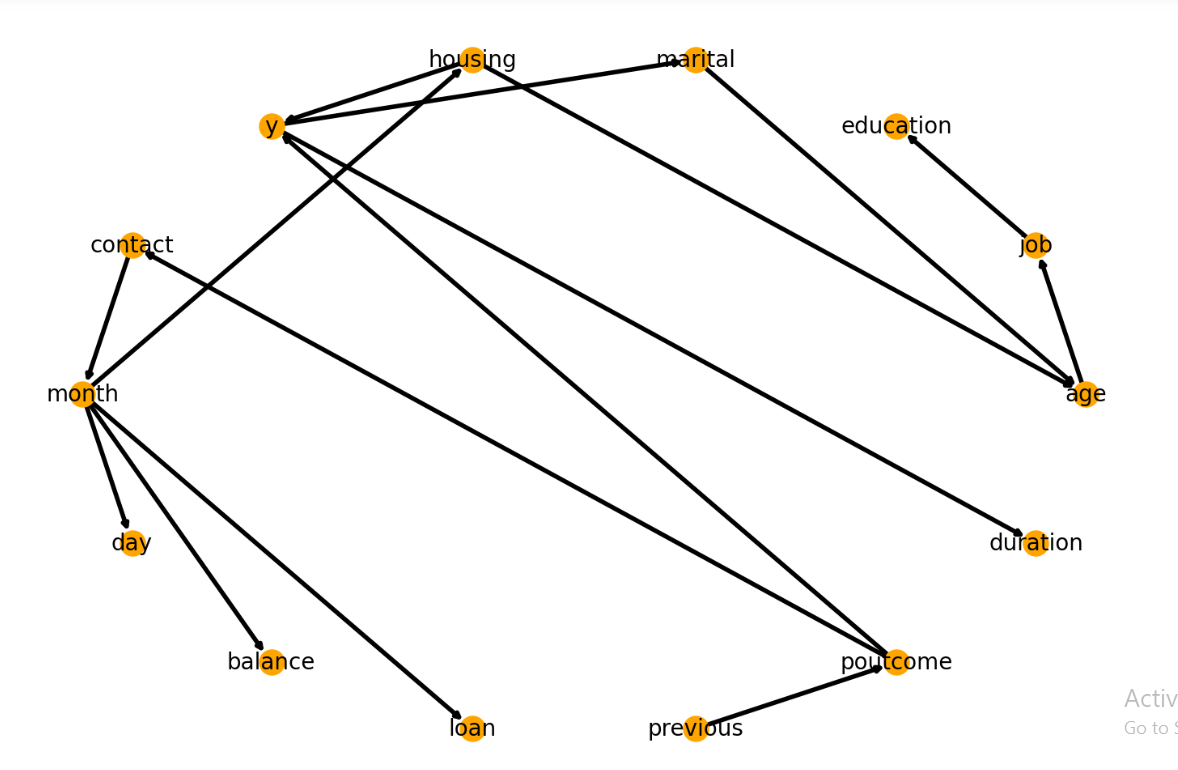
**Approach 2:** Bayesian Network Modelling

The second approach employs a Bayesian Network to address the credit-scoring problem using pgmpy python library. The steps involved are as follows:

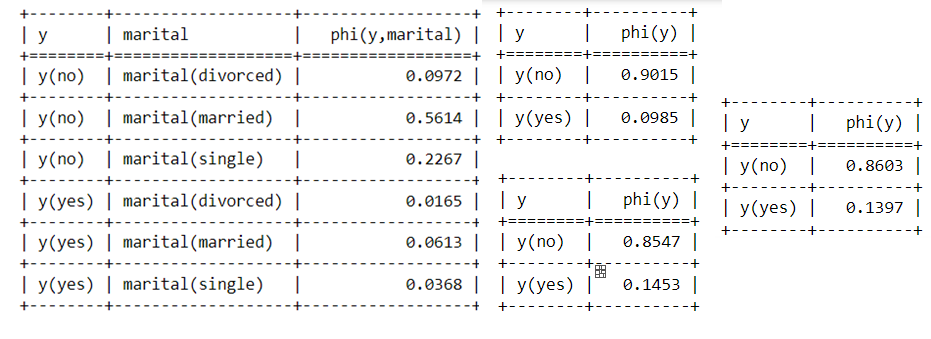
1. Data Preprocessing and Binning: Dropping unnecessary features. Binning of numerical features using equal frequency discretization. Conversion of numerical values into categorical bins for specific features.



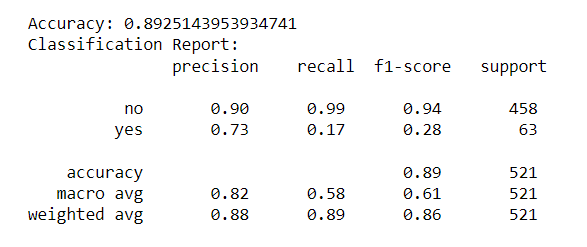
1. Bayesian Network Structure Learning: Application of HillClimbSearch to estimate the network structure. Utilization of BIC scoring method to determine the optimal structure. Visualization of the Bayesian Network using networkx and pylab libraries.



1. Model Training and Inference: Fitting the Bayesian Network to the training data using Maximum Likelihood Estimation. Utilization of Variable Elimination for inference on the trained model. Querying the model to obtain probabilistic predictions based on different scenarios.

*  

1. Model Evaluation: Making predictions on the test data using the Bayesian Network model. Calculation of accuracy and generation of a classification report.



**Comparison and Observations**

Complexity:

The Logistic Regression approach involves a relatively simpler modeling technique compared to Bayesian Networks.

Bayesian Networks are capable of representing complex dependencies between variables, making them suitable for capturing intricate credit scoring patterns.

Interpretability:

Logistic Regression models provide feature importance scores that can be easily interpreted.

Bayesian Networks offer a graphical representation of relationships between variables, aiding in understanding causal dependencies.

Accuracy:

The accuracy of both approaches were around 90% and in both models recall and f1-score were relatively low due to data imbalance in our target feature with denied credit requests significantly higher than accepted ones.

Data Handling:

Bayesian Networks require data preprocessing and discretization to handle categorical and numerical variables effectively.

Logistic Regression may not require as extensive data preprocessing for categorical variables.

**Conclusion:**

In conclusion, both approaches have their merits and drawbacks. In contrast to the Bayesian network approach, the logistic regression approach emphasizes simplicity and the importance of the analysis of features, while the Bayesian network approach is particularly good at modelling complex dependencies and providing graphical insight into variable relationships. The choice between these approaches is dependent upon the particular needs of a credit-scoring problem, on available data and on the required level of interpretability and accuracy.