# Capstone Group 2 FEB24 batch

# Final report

## 1. Summary of Problem Statement, Data, and Findings

## The objective of this project was to predict whether a loan should be approved or denied based on historical loan data while also exploring customer segmentation through clustering. The dataset contained features such as loan amount, term, state, purpose, and job retention metrics.

## Findings:

## The clustering model identified distinct borrower groups, helping tailor lending strategies.

## Scatter plots revealed significant relationships between loan approval status and other features like loan amounts, state, and purpose. 

## Significant correlation between loan attributes and approval status.

## Identification of critical features affecting approvals.

## A final predictive model was developed with high accuracy, offering reliable insights for loan approval decisions.

## 2. Overview of the Final Process

## Methodology:

## Data Preprocessing:

## Cleaned data by removing irrelevant features and addressing missing values.

## Encoded categorical variables and scaled numerical ones.

## Data Exploration: Assessed distributions, identified missing values, and visualized key relationships.

## Preprocessing: Cleaned and transformed data by removing irrelevant columns, encoding categorical variables, and scaling numerical features.

## Clustering Analysis:

## Applied K-Means clustering to segment borrowers based on loan-related attributes.

## Predictive Modelling:

## Trained models like Logistic Regression, Random Forest, and Gradient Boosting for classification.

## Combination: Used ensemble techniques to improve performance, leveraging strengths of individual models.

## Scatter Plot Analysis:

## Visualized relationships between loan approval status and key numerical features for insights.

## 3. Step-by-Step Walkthrough of the Solution

## Data Analysis:

## Visualized distributions and outliers using histograms and boxplots.

## Analysed categorical features with word clouds and bar charts.

## Data Cleaning:

## Removed redundant columns like retained job.

## Addressed missing values and outliers.

## Feature Engineering:

## Performed chi-square tests for categorical relationships.

## Derived new features based on domain insights.

## Scatter Plot Analysis:

## Plotted SBA loan approvals against numerical features like Loan Amount, Term, and Retained Jobs.

## Observations:

## Higher loan amounts were less likely to be approved.

## Borrowers with retained jobs above a threshold were favoured in approvals.

## Specific states had higher approval rates, reflecting regional economic factors.

## Model Building:

## Trained Logistic Regression, Random Forest, and Gradient Boosting models, followed by hyperparameter tuning.

## Final model: Random Forest with SMOTE to handle class imbalance, achieving high accuracy.

## Clustering Analysis:

## Used the K-Means algorithm with features like Loan Amount, Term, and Approval Fiscal Year.

## Opted for 3 clusters, identified through the elbow method, representing high-risk, medium-risk, and low-risk borrower groups.

## Insights: Borrowers in Cluster 3 (low risk) had higher loan approval rates and lower loan amounts.

## 4. Model Evaluation

## The Random Forest model achieved:

## Accuracy: 98.81%

## Precision: 98.85%

## Recall: 98.81%

## F1 Score: 98.82%

## The Logistic regression model achieved:

## Accuracy: 98.68%

## Precision: 98.68%

## Recall: 98.68%

## F1 Score: 98.68%

## The XGBoost model achieved:

## Accuracy: 98.75%

## Precision: 98.78%

## Recall: 98.75%

## F1 Score: 98.76%

## The clustering model demonstrated distinct segmentation:

## Cluster 1: High-risk borrowers, low approval rates.

## Cluster 2: Medium-risk borrowers with moderate loan amounts.

## Cluster 3: Low-risk borrowers, high approval rates.

## 5. Comparison to Benchmark

## Initial benchmarks were Logistic Regression with an accuracy of 78%. Incorporating Random Forest and Gradient Boosting models with feature engineering and SMOTE improved performance significantly. Clustering added an exploratory dimension by identifying actionable borrower groups.

## 6. Visualizations

## Scatter Plots:

## SBA approvals vs. Loan Amount: Negative trend; higher amounts correlated with fewer approvals.

## SBA approvals vs. Retained Jobs: Positive trend; more retained jobs led to higher approval rates.

## Clustering:

## Visualized clusters using a 2D projection. The segmentation highlighted borrower behaviour and approval tendencies.

## Feature Importance:

## Bar chart showing the top contributors, such as Loan Amount, Term, and Purpose.

## Word Clouds: Illustrated distributions in categorical variables like State and Purpose.

## Boxplots: Showcased outlier treatment in loan amounts.

## Feature Importance Chart: Highlighted the top contributors to loan approval decisions.

## Confusion Matrix: Provided a detailed breakdown of model predictions.

**Heatmap**

A screenshot of a computer

Description automatically generated

**Outliers**

A screenshot of a computer

Description automatically generated

Loan Term (in Month)

A screenshot of a computer

Description automatically generated

Gross Approval amount

A screenshot of a computer

Description automatically generated

Average Loan Approvals per month

A screenshot of a computer

Description automatically generated

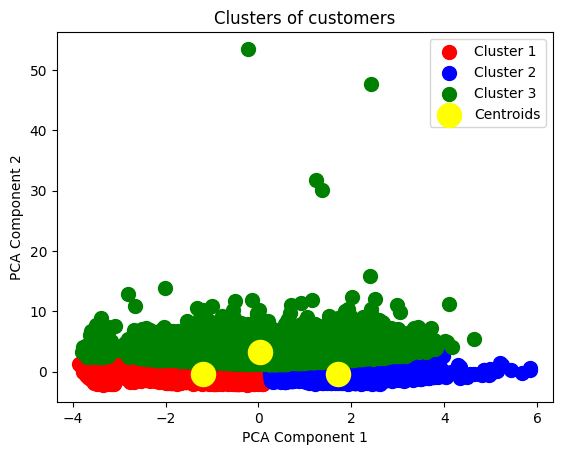
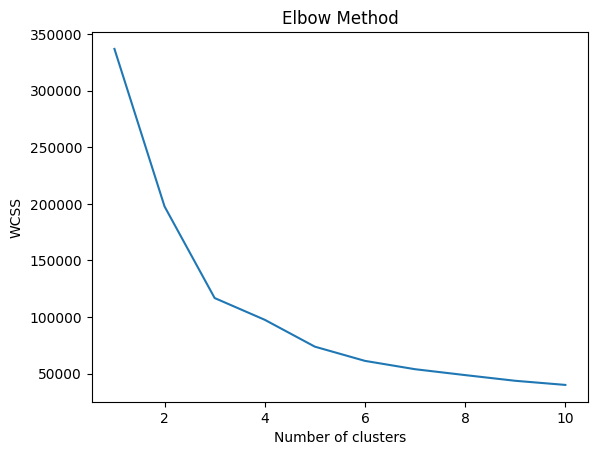
**Non-Default vs Default**

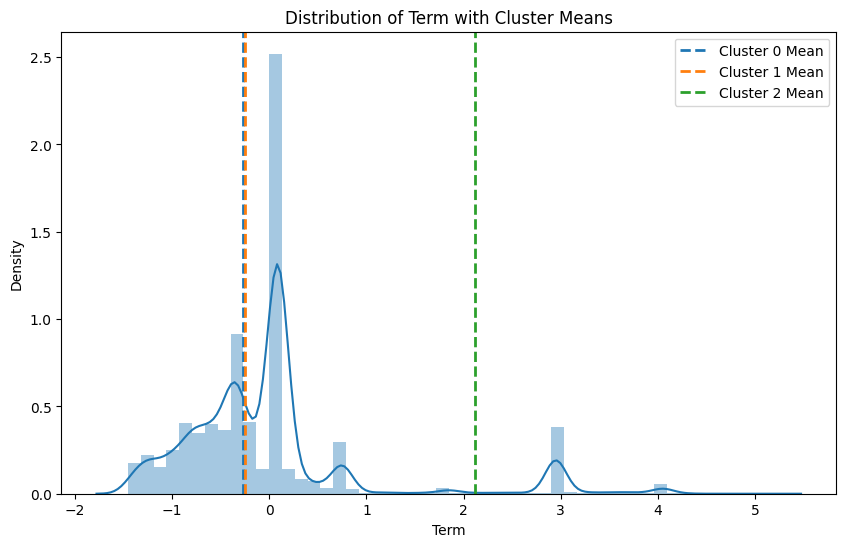
A screenshot of a computer

Description automatically generated

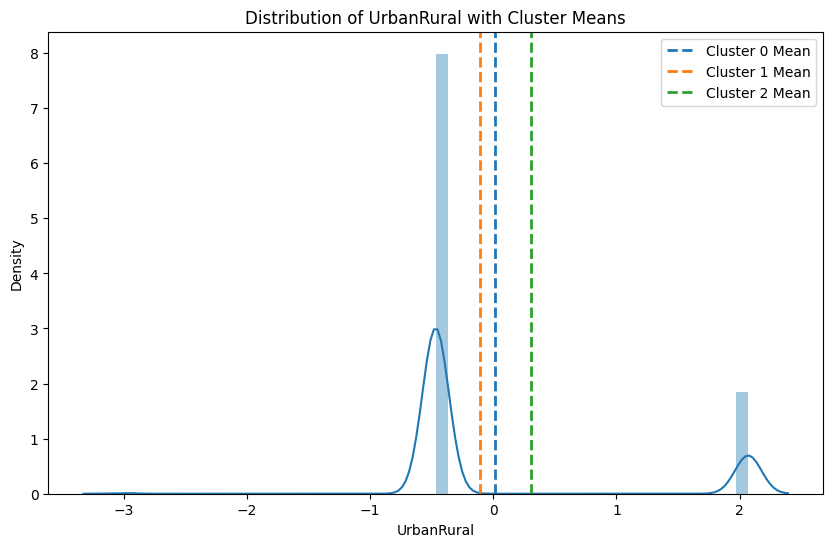
**Cluster Formation distinction**

* 1. **Kmeans Clustering**

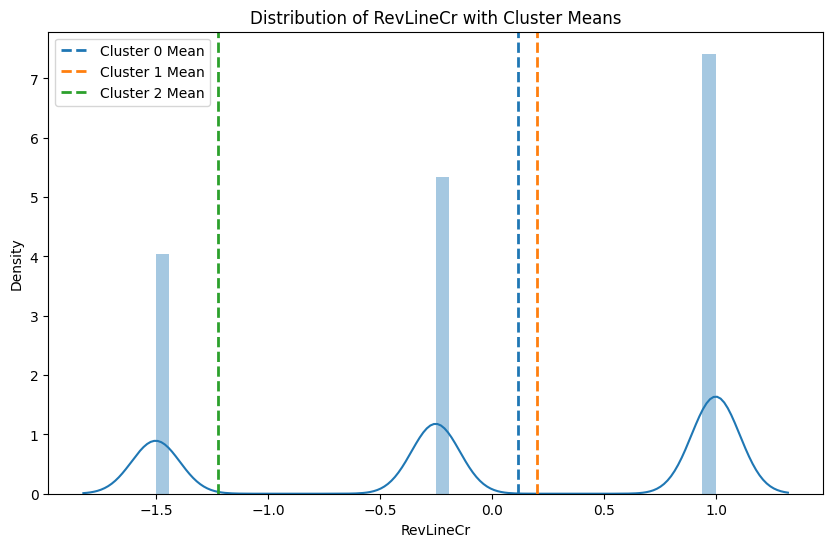


**Distribution of Loan term for different Clusters**

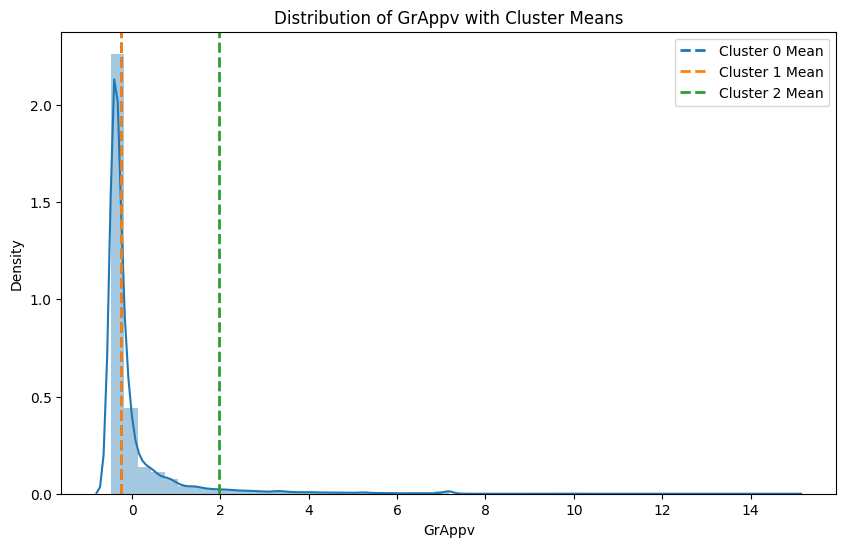
**Distribution of Urban rural Exposure**



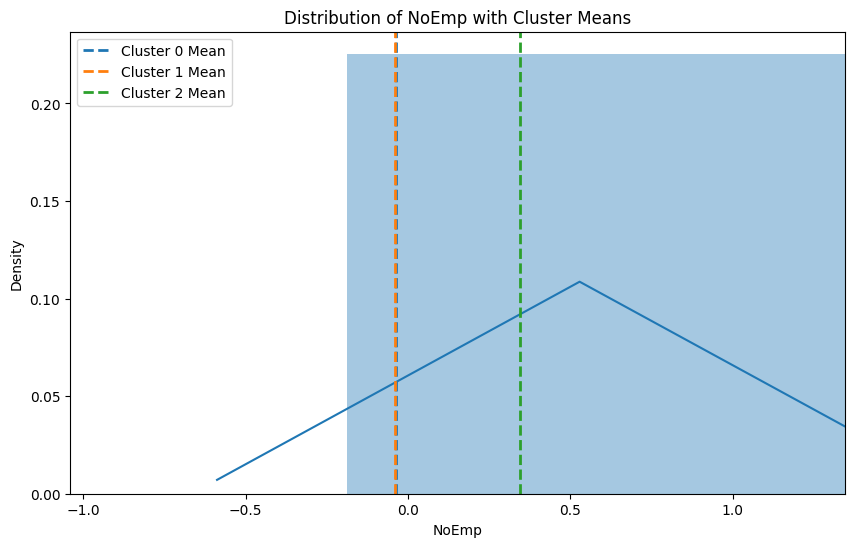
**Distribution of Revolving line of credit**



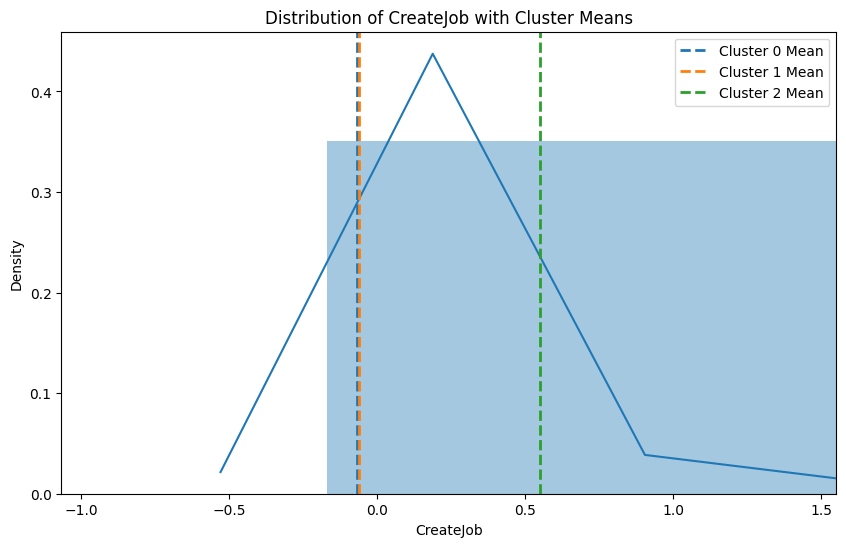
**Distribution of Gross Approval loan Amount**



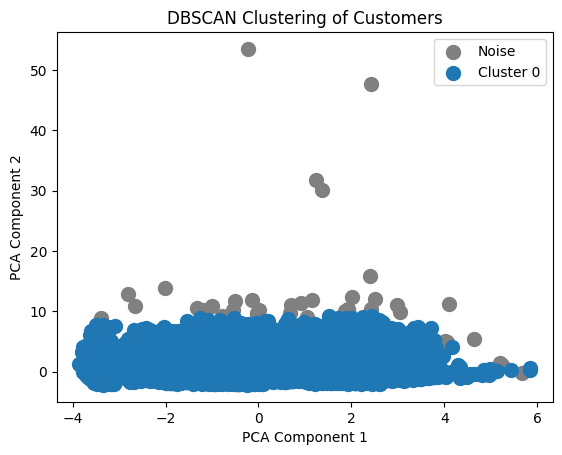
**Distribution of No. of Employees in Company taking Loans**



**Distribution Of jobs created by Companies taking loans**



The insights as follows from the normal distribution graphs made for each feature along with the mean of clusters as vertical lines standing in each feature : The businesses taking higher approved loans were creating more jobs, had more no. Of employees, had a longer loan term duration, had more urban rural exposure, had less no. of revolving line of credit. Whereas on the other hand there are business that has differented by the no. of revolving line of credi they have and the urban rural exposure. This clearly indicates us that there are 3 distinct types of clusters and each have unique feature or features based on which they have been clustered together.

**DBSCAN Clusering model**

## 7. Implications

## The clustering analysis aids lenders in tailoring loan policies for distinct borrower groups. Predictive modelling ensures efficient and consistent loan decisions, reducing human bias and errors. Recommendations:

## Integrate the predictive model and clustering insights into the loan approval workflow.

## Periodically update models with new data for adaptability.

## 8. Limitations

## Clustering assumptions (e.g., linear separability) may not hold in complex datasets.

## Class imbalance and outlier issues may affect prediction performance in unseen data.

## Limited features restrict broader insights; integrating external datasets could enhance the model.

## Overfitting due to high accuracy and precision and recall on train data.

## 9. Closing Reflections

## This project underscored the value of combining predictive and unsupervised learning techniques for comprehensive insights. Key Learnings:

## Effective feature engineering and data visualization are pivotal for impactful solutions.

## Clustering adds a valuable exploratory dimension to predictive modelling.

## Focus more on exploratory data analysis to uncover hidden patterns.

## Future Scope:

## Experiment with advanced clustering methods like DBSCAN for nuanced segmentation.

## Incorporate temporal analysis to understand changing trends in loan approvals.