# Project Report: Loan Approval Analysis and Business Recommendations

---

## Project Overview:

In this project, I analyzed a **loan approval dataset** to understand key factors influencing loan approvals. By examining variables such as **Credit Score**, **Loan Amount**, **Repayment Status**, and **Annual Income**, I aimed to provide actionable recommendations that help improve the **loan approval process** and **customer segmentation** for the financial institution.

## Business Objective:

The goal is to optimize the loan approval process, reduce the risk of defaults, and increase the number of approved loans for qualified customers.

- **Increase loan approvals** for eligible customers.
- **Identify high-risk customers** to minimize defaults.
- **Improve customer satisfaction** by offering tailored loan products.

## Key Data Insights:

After analyzing the data, we found several patterns that can help the business make data-driven decisions:

1. **Credit Score and Loan Approval:**

    - Customers with higher credit scores are more likely to have their loans approved. Customers with a **credit score above 650** have a higher probability of being approved.

**Business Impact**: The financial institution may want to focus on marketing premium products or services for customers with higher credit scores, offering more favorable loan terms.

2. **Loan Amount and Approval Rate:**
    - Loans with amounts **below $30,000** have a significantly higher approval rate compared to loans with amounts above $50,000. This could be due to the perceived risk of larger loans.

**Business Impact**: The company could consider increasing the loan approval threshold for smaller amounts to reach a larger customer base. For larger loans, additional verification processes may be required.

3. **Repayment Status and Loan Defaults:**

    - A key variable influencing loan approval is **repayment behavior**. Customers with a history of late repayments have a significantly higher default rate.

**Business Impact**: The institution should improve its credit risk assessment by prioritizing customers with a clean repayment history. It may also consider offering **financial education programs** to customers at risk of defaulting.

4. **Income and Loan Approval:**

    - Customers with **higher income levels** (above $75,000) tend to be more likely to be approved for loans.

**Business Impact**: The bank can create **personalized loan offers** for higher-income groups, potentially offering larger loans at more competitive interest rates. However, they should not ignore middle- or lower-income applicants, as these segments also contribute to the overall portfolio.

5. **Geographic Trends:**

    - In certain regions, loans are more likely to be approved due to local economic conditions or demand for specific financial products.

**Business Impact**: The bank can **tailor marketing campaigns** to specific regions where loan approval rates are higher and adjust loan products for regions with lower approval rates.

---

## Business Recommendations:

**1. Improve Customer Segmentation:**

- **Actionable Insight**: Segment customers based on their **credit scores**, **loan amounts**, and **income levels** to offer personalized loan products. This allows the bank to target marketing and offers to the right customer segments.

- **Recommendation**: Develop segmented loan products for **high-income** customers, **mid-income** customers, and **low-income** customers. Use credit scores to differentiate the risk and offer tailored interest rates.

**2. Reduce Risk of Loan Defaults:**

- **Actionable Insight**: Customers with a history of late repayments pose a higher risk of default. The current approval process might not be fully capturing this risk.

- **Recommendation**: Implement a **robust risk assessment model** that takes into account **repayment behavior** alongside credit scores. Offer customers with poor repayment history financial literacy programs or **pre-approved smaller loans** to rebuild trust.

**3. Increase Loan Approval Rate for Smaller Loans:**

- **Actionable Insight**: Loans under $30,000 have a significantly higher approval rate compared to larger loans. This could be due to less perceived risk for smaller loans.

- **Recommendation**: The bank should consider offering **faster approval processes** for smaller loans (e.g., loans under $30,000) to increase customer engagement. For larger loans, more stringent criteria can be implemented, but with a streamlined process for qualifying applicants.

**4. Tailor Loan Offers Based on Income:**

- **Actionable Insight**: Higher-income customers are more likely to be approved, but middle- or lower-income groups should not be neglected.
- **Recommendation**: Create **tailored loan products** for each income group. For example, offer **low-interest loans** for higher-income customers, while providing **affordable, flexible repayment terms** for middle- and lower-income customers.

**5. Target Geographic Areas with High Approval Rates:**

- **Actionable Insight**: Loans in certain regions show higher approval rates due to economic conditions.
- **Recommendation**: Use **geographic data** to optimize the **targeting of loan products**. Market loan products that are tailored to specific regional needs, and offer additional incentives or lower rates in regions with lower approval rates to drive more loan applications.

---

## Potential Business Impact of Implementing Recommendations:

By adopting these recommendations, the business could see:

- **Increased Loan Volume**: By approving more small loans and offering tailored products to high-income customers, the bank can increase loan volume.
- **Reduced Risk**: By improving credit risk assessment and offering financial education programs, the bank can reduce the risk of defaults.
- **Higher Customer Satisfaction**: Personalized loan products and marketing will resonate better with different customer segments, leading to improved customer loyalty and satisfaction.
- **Optimized Loan Portfolio**: A well-balanced portfolio across different loan amounts and income levels will lead to a more robust and sustainable loan portfolio.

---

## Conclusion:

This analysis of loan approval data provides critical insights into how the loan approval process can be optimized from a business perspective. By segmenting customers, reducing risk, and offering tailored loan products, the financial institution can increase loan approval rates, improve customer satisfaction, and minimize defaults.

Code Snippet for the sample data

```python
# 2. Finance Data Analysis (Loan Approval Prediction)
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
from sklearn.preprocessing import StandardScaler
```

```python
# Step 1: Generate synthetic data
np.random.seed(42)
```

```python
# Generating 1000 loan application records
data_size = 1000
```

```python
# Features
credit_score = np.random.randint(600, 800, data_size)
loan_amount = np.random.randint(5000, 50000, data_size)
loan_term = np.random.choice([1, 2, 3, 5, 10], data_size)
repayment_status = np.random.choice(['On Time', 'Late', 'N/A'], data_size)
```

```python
# Loan approval decision (0 = Rejected, 1 = Approved)
# Simulating loan approval logic based on some factors
loan_approval = (credit_score > 650) & (loan_amount < 30000) & (repayment_status != 'Late')
loan_approval = loan_approval.astype(int)  # Convert boolean to 0 or 1
```

```python
# Step 2: Create DataFrame
df = pd.DataFrame({
    'CreditScore': credit_score,
    'LoanAmount': loan_amount,
    'LoanTerm': loan_term,
    'RepaymentStatus': repayment_status,
    'LoanApproval': loan_approval
})
```

```python
# Step 3: Preprocess the data
# Convert 'RepaymentStatus' to numeric (On Time=1, Late=0, N/A=-1 for missing)
df['RepaymentStatus'] = df['RepaymentStatus'].map({'On Time': 1, 'Late': 0, 'N/A': -1})
```

```python
# Drop rows where RepaymentStatus is N/A (optional, based on your requirement)
df = df.dropna(subset=['RepaymentStatus'])
```

```python
# Features (X) and Target (y)
X = df[['CreditScore', 'LoanAmount', 'RepaymentStatus']]
y = df['LoanApproval']
```

```python
# Step 4: Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

```python
# Step 5: Feature Scaling (Important for Logistic Regression)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

```python
# Step 6: Build Logistic Regression Model
model = LogisticRegression(random_state=42)
model.fit(X_train_scaled, y_train)
```

▾ LogisticRegression

▶ Parameters

```python
# Step 7: Evaluate the Model
y_pred = model.predict(X_test_scaled)
```

```python
# Accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f"Model Accuracy: {accuracy * 100:.2f}%")
```

```
Model Accuracy: 78.00%
```

```python
# Classification Report for additional evaluation metrics
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.90      0.85       202
           1       0.72      0.54      0.62        98

    accuracy                           0.78       300
   macro avg       0.76      0.72      0.73       300
weighted avg       0.77      0.78      0.77       300
```

```python
# Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
print("\nConfusion Matrix:")
print(cm)
```

```
Confusion Matrix:
[[181  21]
 [ 45  53]]
```

## 1. Classification Report

The **Classification Report** provides the following metrics for both classes (`0` = Rejected, `1` = Approved):

- **Precision**: The proportion of positive predictions (approvals) that were actually correct.

  Formula:

  $$\text{Precision} = \frac{TP}{TP + FP}$$

  Where:

  - **TP** = True Positives (correct predictions of loan approval)
  - **FP** = False Positives (incorrect predictions of loan approval)
- **Recall**: The proportion of actual positives (approved loans) that were correctly predicted by the model.

  Formula:

  $$\text{Recall} = \frac{TP}{TP + FN}$$

  Where:

  - **FN** = False Negatives (loan approvals that the model failed to predict)
- **F1-Score**: The harmonic mean of precision and recall. This gives a balance between both metrics, and is especially useful when the dataset is imbalanced.

Formula:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **Support**: The number of actual occurrences of the class in the test set.

Here's the breakdown for each class in the report:

```
[[181  21]  # True Negatives (TN) and False Positives (FP)
 [ 45  53]]  # False Negatives (FN) and True Positives (TP)
```

- **True Negatives (TN)**: 181 — The number of correctly predicted `Rejected` loans.
- **False Positives (FP)**: 21 — The number of `Rejected` loans incorrectly predicted as `Approved`.
- **False Negatives (FN)**: 45 — The number of `Approved` loans incorrectly predicted as `Rejected`.
- **True Positives (TP)**: 53 — The number of correctly predicted `Approved` loans.

**For `Loan Rejected` class (label `0`):**

For `Loan Rejected` **class (**`label 0`**):**

- **Precision for Class 0 (Rejected)**:

  Precision tells you how many of the `Rejected` loans predicted by the model were actually correct.

$$\text{Precision for Class 0} = \frac{TN}{TN + FP} = \frac{181}{181 + 21} = \frac{181}{202} = 0.896$$

- **Precision**: 0.896 — Of all the instances predicted as `0` (Rejected), 90% were actually rejected loans.

$$\text{Recall for Class 0} = \frac{TN}{TN + FN} = \frac{181}{181 + 45} = \frac{181}{226} = 0.80$$

- **Recall**: 0.80 — The model correctly identified 80% of all the actual rejected loans.
- **F1-Score**: 0.85 — This is a good score, as it's high enough to indicate a solid performance on rejecting loans.

**For `Loan Approved` class (label `1`):**

For `Loan Approved` class (`label 1`):

- **Precision for Class 1 (Approved)**:

  Precision tells you how many of the `Approved` loans predicted by the model were actually correct.

$$\text{Precision for Class 1} = \frac{TP}{TP + FP} = \frac{53}{53 + 21} = \frac{53}{74} = 0.716$$

- **Precision**: 0.72 — Of all the instances predicted as `1` (Approved), 72% were actually approved loans.

**Recall for Class 1 (Approved)**:

Recall tells you how many of the actual `Approved` loans were correctly predicted by the model.

$$\text{Recall for Class 1} = \frac{TP}{TP + FN} = \frac{53}{53 + 45} = \frac{53}{98} = 0.540$$

- **Recall**: 0.54 — The model only identified 54% of the actual loan approvals. This indicates that the model is missing many approved loans.

**F1-Score Calculation:**

The **F1-Score** is the harmonic mean of **Precision** and **Recall**, and it gives a balanced measure of the two.

The formula is:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

**For Class 0 (Rejected):**

Using the calculated precision and recall:

$$\text{F1-Score for Class 0} = 2 \times \frac{0.80 \times 0.90}{0.80 + 0.90} = 2 \times \frac{0.72}{1.70} = 0.847$$

(Rounded to 0.85 in the classification report.)

**For Class 1 (Approved):**

Using the calculated precision and recall:

$$\text{F1-Score for Class 1} = 2 \times \frac{0.72 \times 0.54}{0.72 + 0.54} = 2 \times \frac{0.3888}{1.26} = 0.619$$

## 2. Accuracy

- **Accuracy**: 0.78 or 78%. This means that, overall, the model is correct 78% of the time across all instances.