**FINANCIAL ANALYSIS FOR LOAN APPROVAL**

**Quick Summary:**

**Objective:** I worked on building a model that helps us predict whether a loan application will be **approved** or **denied**. This will help us better understand the factors that affect loan decisions and improve the efficiency and accuracy of our loan approval process.

**Step-by-Step Overview:**

1. **Data Preparation:**
   * I started with a dataset containing information about loan applications, including variables such as **Credit Score**, **Annual Income**, **Loan Amount**, **Employment Status**, and whether the loan was ultimately **Approved** or **Denied**.
   * I cleaned the data by converting date-related fields into a usable format and created new features like **Employment Length** (how long the applicant has been employed).
   * Additionally, I calculated important financial ratios like:
     + **Debt-to-Income**: The ratio of the applicant's debt to their income.
     + **Loan-to-Income**: How much of the applicant's income is tied to the loan amount they applied for.
     + **Debt-to-Loan**: The ratio of outstanding debt to the loan amount.
2. **Exploratory Analysis:**
   * I explored the data visually to identify trends and patterns. For example:
     + **Credit Score**: Applicants with higher credit scores were more likely to be approved.
     + **Employment Status**: Those who were **employed** were more likely to be approved compared to those who were **self-employed** or **unemployed**.
     + I also plotted relationships between variables like **Annual Income** and **Loan Amount**, showing how they vary based on whether a loan was approved or denied.
3. **Model Building:**
   * I used **Logistic Regression** as the primary model. Logistic regression is a method that helps predict whether something will happen based on several factors.
     + I split the data into a training set (for learning patterns) and a test set (for evaluating the model’s performance).
     + The model looked at all the factors (like credit score, income, debt, etc.) to estimate the chances of the loan being approved or denied.
4. **Model Performance:**
   * The model's **accuracy** was about **51%**, which means it predicted loan outcomes correctly a little more than half of the time. This accuracy is just above what we'd get by randomly guessing, so there’s room for improvement.
   * Other metrics like **precision** and **recall** helped us understand how well the model predicted approved and denied loans. For example:
     + **Precision** tells us that when the model predicted a loan would be approved, it was right about **51%** of the time.
     + **Recall** tells us that the model correctly identified **36%** of all approved loans.
5. **Tuning and Comparison:**
   * I also tried improving the model by adjusting certain settings (hyperparameters), but the results didn’t improve much. Additionally, I compared Logistic Regression with other models like **Decision Trees** and **Random Forests**, but their performance was similar or slightly lower.
6. **Key Insights:**
   * Factors like **Credit Score**, **Annual Income**, and **Employment Status** played a significant role in predicting loan outcomes.
   * Applicants with **higher credit scores** and **higher income** were more likely to have their loans approved, while those with high **debt** or **low income** were more likely to be denied.

**Deep Dive Into Modelling Result**

After deep dive into trying various models for this dataset, lets check the outcome of the result and the reason why logistic regression is chossing to built the final model even though it performed slightly better than random guessing and here’s a quick comparison of their performances:

1. **Logistic Regression**:
   * **Initial model**:
     + Accuracy: 0.5100
     + ROC AUC Score: 0.5086
     + Precision: 0.5143
     + Recall: 0.3600
     + F1 Score: 0.4235
   * **After tuning**:
     + Accuracy: 0.4600
     + ROC AUC Score: 0.4830
     + Precision: 0.4231
     + Recall: 0.2200
     + F1 Score: 0.2895
   * This model's performance dropped after tuning, particularly in recall and F1-score.
2. **Decision Tree**:
   * Training Accuracy: 0.515
   * Test Accuracy: 0.51
   * After tuning, test accuracy dropped to 0.4950. This model also shows modest performance, close to the baseline accuracy.
3. **Random Forest**:
   * Training Accuracy: 1.0 (indicating **overfitting**)
   * Test Accuracy: 0.47 (indicating a drop in generalization ability)
   * After tuning, accuracy remained at 0.47. Though its precision and recall are closer to each other than the previous models, overfitting is a significant issue.
4. **XGBoost**:
   * Initial Accuracy: 0.4900
   * ROC AUC Score: 0.4830
   * Precision: 0.4889
   * Recall: 0.4400
   * F1 Score: 0.4632
   * After tuning, its cross-validation score improved to 0.4712, but its performance on the test set was still comparable to other models (accuracy around 0.47).

**Observation:**

I laterrealized that none of the models show standout performance, with accuracy and other metrics hovering around the baseline (~0.5). However, **Logistic Regression** initially performed better than the others with an accuracy of **0.51** and a relatively decent F1-score, but after tuning, performance dropped.

**Which of the model should be the best for final modelling?:**

It depends!

Let’s take a closer look at the key performance metrics for both models to help clarify which one performs better.

**Decision Tree:**

* **Training Accuracy**: 0.515
* **Test Accuracy**: 0.510
* **Confusion Matrix:** [[71, 29], [69, 31]]
* **Precision**:
  + Class 0 (Denied): 0.49
  + Class 1 (Approved): 0.52
* **Recall**:
  + Class 0: 0.71
  + Class 1: 0.31
* **F1-Score**:
  + Class 0: 0.58
  + Class 1: 0.39

**Logistic Regression (initial model):**

* **Test Accuracy**: 0.5100
* **ROC AUC Score**: 0.5086
* **Precision**: 0.5143
* **Recall**: 0.3600
* **F1-Score**: 0.4235
* **Confusion Matrix**: [[66, 34], [64, 36]]

**Comparison:**

1. **Accuracy**: Both models have the same test accuracy (0.51), so purely based on accuracy, they perform equally.
2. **Recall**:
   * The **Decision Tree** has a much better recall for the negative class (Denied) at 0.71, meaning it correctly identifies most denied loans. However, its recall for the positive class (Approved) is low at 0.31.
   * **Logistic Regression** has a more balanced but lower recall for both classes: 0.66 for Denied and 0.36 for Approved.
3. **F1-Score**:
   * For the **Decision Tree**, the F1-score for the Denied class (0.58) is higher than Logistic Regression's (0.57), but its F1-score for the Approved class (0.39) is lower than Logistic Regression (0.42).
4. **Confusion Matrix**:
   * Both models have very similar confusion matrices, with around 66/71 true negatives (correctly predicted Denied loans) and around 36/31 true positives (correctly predicted Approved loans). However, the Decision Tree is more aggressive in predicting Denied loans, resulting in higher true negatives but lower true positives.

**Final Thoughts:**

The **Decision Tree** seems better suited if the cost of **false positives (wrongly approved loans)** is high, since it’s better at identifying denied loans (higher recall for Class 0). However, if you need more balanced performance between both classes (e.g., both false positives and false negatives are equally important), **Logistic Regression** might still be the better choice.

**Decision:**

Logistic regression for balanced predictions

**Insight into Risk factors affecting loan outcomes**

To provide insights into the **risk factors affecting loan outcomes** using the **Logistic Regression** model, we need to analyze the **coefficients** of the trained model. These coefficients help us understand how each feature influences the likelihood of a loan being approved or denied. Here’s how we can interpret the results:

**Step 1: Review the Coefficients**

The coefficients of the Logistic Regression model represent how each feature impacts the log-odds of the outcome (in this case, loan approval or denial). Here’s a breakdown of what we can learn:

* **Positive Coefficients**: These increase the likelihood of the loan being approved (Class 1).
* **Negative Coefficients**: These increase the likelihood of the loan being denied (Class 0).

**Step 2: Interpretation of Key Risk Factors**

Let’s now focus on interpreting the features and their coefficients:

**1. Credit Score (Positive Coefficient):**

* If **Credit Score** has a **positive coefficient**, it means that as the credit score increases, the likelihood of loan approval increases. Borrowers with higher credit scores are seen as less risky, so they are more likely to get approved.

**2. Annual Income (Positive Coefficient):**

* A **positive coefficient** for **Annual Income** means that higher-income applicants are more likely to get approved for loans. This makes sense since lenders prefer borrowers with higher income, as they are perceived to have more capacity to repay loans.

**3. Outstanding Debt (Negative Coefficient):**

* If **Outstanding Debt** has a **negative coefficient**, it indicates that higher levels of existing debt reduce the likelihood of loan approval. This suggests that applicants with more outstanding debts are considered riskier because they may struggle to manage additional financial obligations.

**4. Employment Status (Categorical):**

* For categorical variables like **Employment Status** (with one-hot encoding for each category), each category (e.g., **Employed**, **Self-Employed**, **Unemployed**) will have its own coefficient.
  + If the coefficient for **Unemployed** is negative, it means that being unemployed decreases the chances of loan approval.
  + **Self-Employed** status might also have a lower coefficient compared to **Employed**, indicating slightly higher risk in lending to self-employed applicants.

**5. Debt-to-Income Ratio (Negative Coefficient):**

* A **negative coefficient** for **Debt-to-Income Ratio** means that as this ratio increases, the likelihood of loan denial also increases. This is intuitive because a higher debt-to-income ratio implies that a borrower has significant debt relative to their income, which increases the risk of loan default.

**6. Loan-to-Income Ratio (Negative Coefficient):**

* A **negative coefficient** for the **Loan-to-Income Ratio** suggests that if the requested loan amount is a large percentage of the applicant's income, it decreases the likelihood of approval. High loan-to-income ratios suggest that the borrower is stretching their finances too thin.

**Step 4: Insights from Coefficient Magnitudes**

* **The larger the absolute value of the coefficient**, the more important that feature is in determining the outcome. Features with smaller absolute coefficients have less influence.
* A **positive sign** indicates that the feature increases the likelihood of approval, while a **negative sign** decreases it.

For example, if **Credit Score** has the largest positive coefficient, it might be the most critical factor in determining whether a loan is approved. Conversely, if **Outstanding Debt** has the largest negative coefficient, it would be the most important risk factor contributing to loan denial.

**Conclusion: Key Risk Factors**

* **Credit Score** and **Annual Income** are usually positive predictors, indicating that higher values increase the likelihood of loan approval.
* **Outstanding Debt**, **Debt-to-Income Ratio**, and **Loan-to-Income Ratio** are typically **negative** predictors, meaning that higher values of these features contribute to a higher risk of loan denial.
* **Employment Status** (especially if Unemployed or Self-Employed) can also play a significant role in determining loan outcomes.

By interpreting the coefficients from the logistic regression model, you can effectively identify the risk factors that are most influential in determining whether a loan is approved or denied.

**Recommendations for better performing model:**

While the model gives us some insight, we can further improve it by looking at more advanced techniques, **gathering additional data**, or fine-tuning the model for better predictions. This way, we can better understand the risks associated with loan approvals and make our process more efficient.