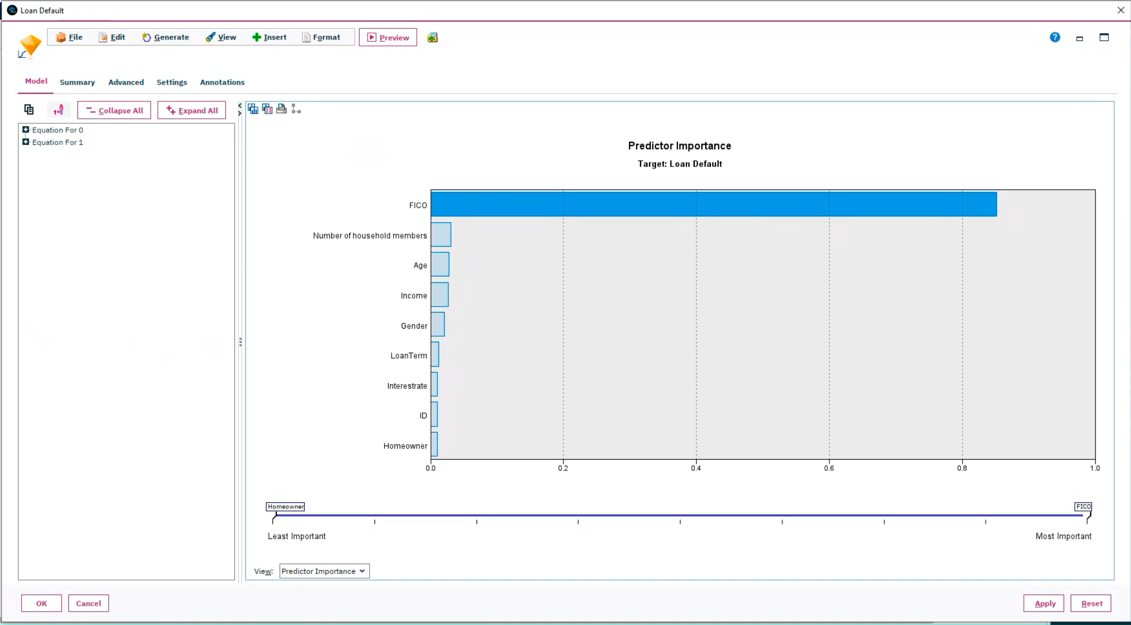
**MKT 568 -Assignment 3**

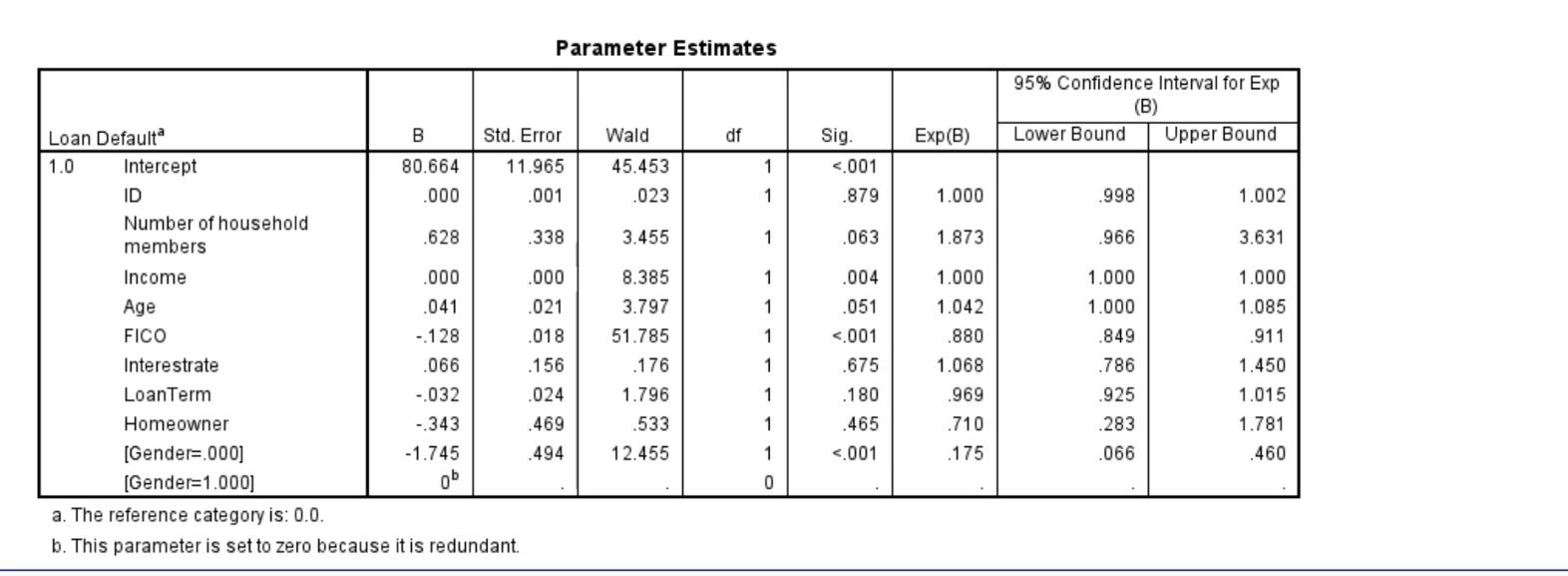
**Names: Noopura Vaidya and Ajinkya Ghorpade**

1. **a.**

A diagram of a diagram

Description automatically generated

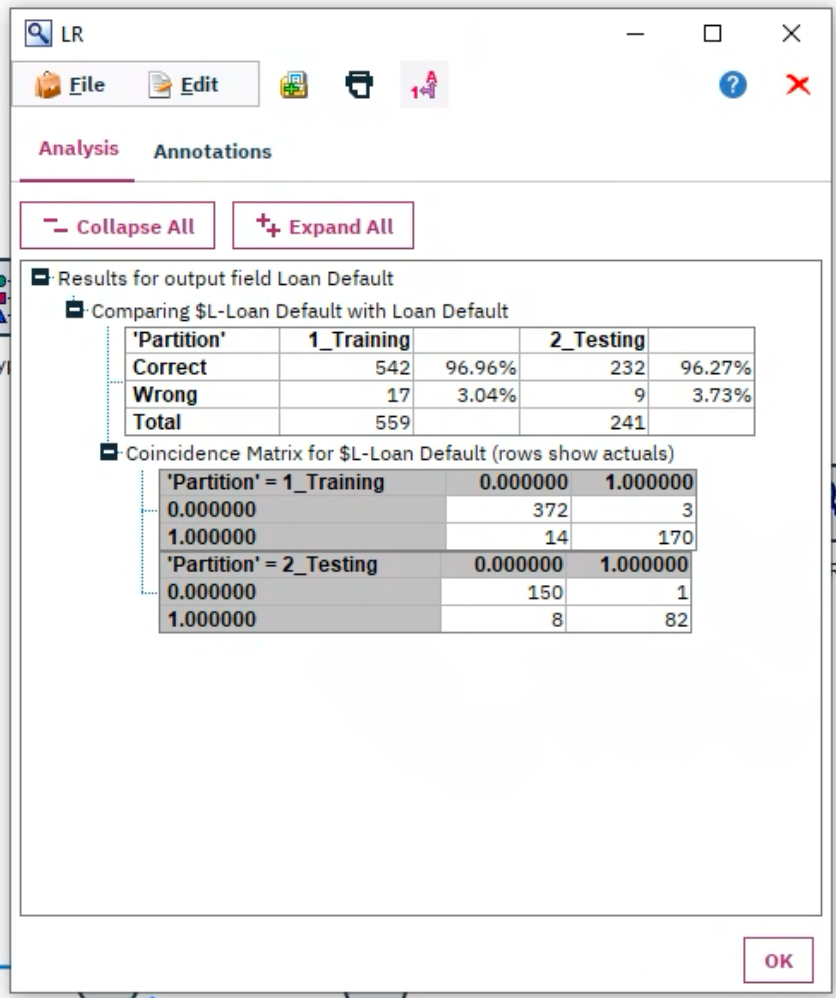


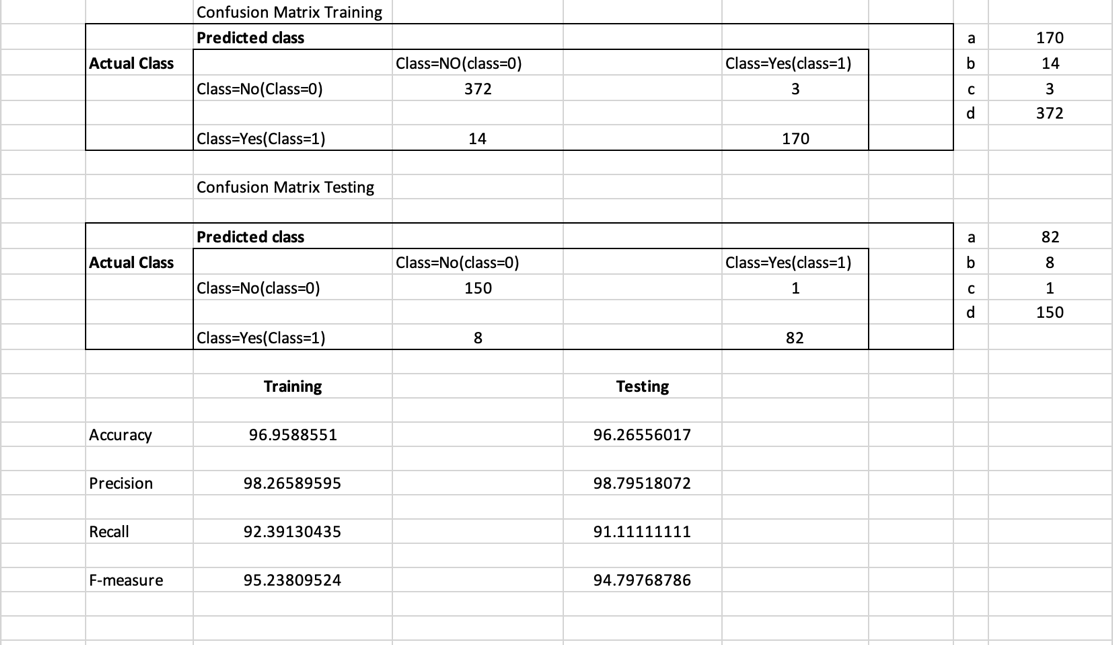


Based on the Predictor Importance chart after we run the logistic regression model, we can deduce which variables have a statistically significant effect on the probability of a customer defaulting on a loan. Here is the interpretation:

* **FICO Score**: This variable has the greatest impact on the default probability. A higher FICO score usually indicates a lower risk of default, while a lower FICO score suggests a higher risk.
* **Number of Households**: The second most significant variable. The number of households could relate to financial stability or obligations.
* **Income**: Third in importance, income is a direct measure of an individual's ability to repay a loan. Higher income decreases the risk of default.
* **Age**: The age of the borrower is also equal to the importance of income as we can reflect financial maturity and stability, which can influence the likelihood of default.
* **Gender**: Gender seems to have a smaller effect compared to other variables. Its significance would be based on statistical findings rather than causation.
* **LoanTerm**: The term of the loan may influence the likelihood of default, with longer terms potentially increasing risk due to the extended commitment but it also has smaller effect.
* **Interest rate:** It is indeed a crucial factor in lending and borrowing as it can significantly impact the borrower's ability to repay the loan. Higher interest rates can increase the cost of borrowing, potentially raising the risk of default if the borrower has a tight budget or fluctuating income. Conversely, lower interest rates might reduce the likelihood of default by making loan payments more affordable. But it does not have importance accoring to the predictor importance chart of this model.
* **ID:** In the context of a predictive model for loan defaults, the ID would not be used as a predictive variable because it does not carry any intrinsic predictive power regarding the behavior of the borrower or the characteristics of the loan. Instead, it is just a label used to distinguish one loan or customer from another.
* **Homeowner**: This indicates whether the customer owns a home, which might affect their financial stability and risk of default.

**b.**





**Formulas used:**

**Accuracy:**  TP+TN / TP+TN+FP+FN

**Precision:** TP/TP+FP

**Recall:** TP/TP+FN

**F-measure:** 2 Precision Recall/ Precision + Recall

**Training Set Performance:**

* **Accuracy**: 96.96% - This suggests the model correctly predicted the loan default status for a high percentage of the training data.
* **Precision**: 98.27% - Indicates that out of all the loans predicted to default, the vast majority did indeed default. This is a high level of precision, showing the model is good at predicting actual defaults.
* **Recall**: 92.39% - This is the proportion of actual defaults that were correctly identified by the model. A recall rate above 90% is considered very good.
* **F-Measure**: 95.24% - This is a measure that balances precision and recall, and a high value here means the model performs well on both fronts.

**Testing Set Performance:**

* **Accuracy**: 96.27% - Close to the training accuracy, suggesting the model generalizes well.
* **Precision**: 98.80% - Even higher than the training precision, indicating excellent performance on the testing data.
* **Recall**: 91.11% - A slight decrease from the training recall but still very high, showing the model remains effective at detecting defaults.
* **F-Measure**: 94.80% - Though slightly lower than in the training set, it's still a high score, reflecting a good balance of precision and recall.

**Confusion Matrix Interpretation:**

**For the Training Set:**

* **True Negatives (TN):** 372 cases were correctly predicted as non-default.
* **False Negatives (FN):** 14 cases were actual defaults but were incorrectly predicted as non-default.
* **False Positives (FP):** 3 cases were actual non-defaults but were incorrectly predicted as defaults.
* **True Positives (TP):** 170 cases were correctly predicted as defaults.

**For the Testing Set:**

* **True Negatives (TN):** 150 cases were correctly predicted as non-default.
* **False Negatives (FN):** 8 cases were actual defaults but were incorrectly predicted as non-default.
* **False Positives (FP):** 1 case was an actual non-default but was incorrectly predicted as a default.
* **True Positives (TP):** 82 cases were correctly predicted as defaults.

**Evaluation Based on the Confusion Matrix:**

* High TN and TP values indicate a high number of correct predictions.
* Low FN and FP values indicate a small number of incorrect predictions.
* In both the training and testing sets, the number of True Positives and True Negatives is high compared to False Positives and False Negatives, which indicates a good performance of the model.
* The model has high accuracy (correct predictions / total predictions), high precision (true positives / predicted positives), and high recall (true positives / actual positives), which are desirable in a good predictive model.

The confusion matrix thus confirms that the model is performing well both in distinguishing the non-default cases (high TN) and in identifying the default cases (high TP). It is also making very few errors in both the training and testing phases, which is shown by the low FP and FN values.

Overall, this model is performing very well and can be **considered a good fit for predicting loan defaults.**

**c.**

A screenshot of a computer

Description automatically generated

* In the logistic regression results provided, gender has indeed been broken down into two categories. Typically, logistic regression requires numerical input for all variables, so categorical variables like gender, which often include categories like 'male' and 'female,' must be encoded numerically. This process is called dummy coding or one-hot encoding.
* Here's a brief explanation of how this works and why only one gender category might have a significant coefficient:
* **Dummy Coding**: The categorical variable 'Gender' is split into two variables, often 'Gender=Male' and 'Gender=Female'. However, only one of these is necessary to capture the effect of gender in a binary classification since the other category is implied by the absence of the first. This avoids multicollinearity issues, where the predictor variables are highly correlated.
* **Reference Category**: By convention, one category is omitted and used as the reference category. In the output, 'Gender=0' likely represents the reference category (which could be male if coded as '0'), and the coefficient for 'Gender' in the model would then represent the change in the log odds of default for the non-reference category (which could be female if coded as '1') compared to the reference category.
* **Significance of the Coefficient**: In the output, the coefficient for 'Gender' is significant if it substantially differs from zero. If only 'Gender=0' is shown with a coefficient, it might be due to a few reasons:
  + The software might automatically exclude the reference category from the output, showing only the category that is being compared to the reference.
  + The coefficient for 'Gender=0' represents the effect of being in that category as opposed to the reference category, and the significance indicates that being in that category (whatever gender it represents) has a statistically significant impact on the likelihood of defaulting on a loan.
* If 'Gender=0' has a significant coefficient, it suggests that there is a significant difference in the likelihood of default between the gender coded as '0' and the gender coded as '1'. If 'Gender=0' is negative, it indicates that being in the '0' category (possibly male) is associated with a lower log-odds of defaulting on a loan compared to being in the '1' category (possibly female), or vice versa if the coefficient is positive.

**d.**

A table with numbers and a few words

Description automatically generated with medium confidence

* The results indicate that one gender is associated with higher odds of being a defaulter. This does not necessarily mean that the model is discriminating, but rather it reflects the patterns in the historical data.
* However, if we use this model for making decisions, it may lead to discrimination against a specific gender. This is an important ethical consideration, and alternative modeling approaches or additional context could be needed to ensure fairness.

In the logistic regression output, we can see the parameter estimates for the model predicting loan default. Specifically, looking at the gender variable, we see two entries: [Gender=0.000] and [Gender=1.000].

The coefficient for [Gender=0.000]is -1.732 with a statistically significant p-value (<.001), which suggests that the odds of defaulting on a loan are lower for the reference group (males if 0 represents males) compared to the base category. Since this is a logistic regression, the actual odds ratio is the exponential of the B coefficient. An odds ratio less than 1 (Exp(B) for [Gender=0.000] is 0.177) indicates that the odds of the event (defaulting) are lower for the reference group than for the comparison group (females if 1 represents females).

The entry for [Gender=1.000] shows a coefficient of 0 by default because it is redundant to model both categories of a binary variable. The model uses one category as the reference (in this case, [Gender=1.000] is the reference category) and compares all other categories against this reference.

To clarify, a positive coefficient would indicate higher odds of defaulting for the category in question relative to the reference category, while a negative coefficient (as we see here for [Gender=0.000] indicates lower odds of defaulting relative to the reference category.

This means that in the data used to train this model, the group coded as 1 (females) has higher odds of defaulting on a loan compared to the group coded as 0 (males), when controlling for other variables in the model.

Now, regarding discrimination, this result by itself does not mean the model is unfair or discriminatory. It is simply capturing the patterns in the data provided. However, if the model were to be used for decision-making without considering the broader context, it might result in decisions that unfairly disadvantage one gender. This is why it is essential to carefully consider how such models are used, and to potentially adjust the model or its application to ensure fairness and compliance with ethical standards and regulations.

**2.**

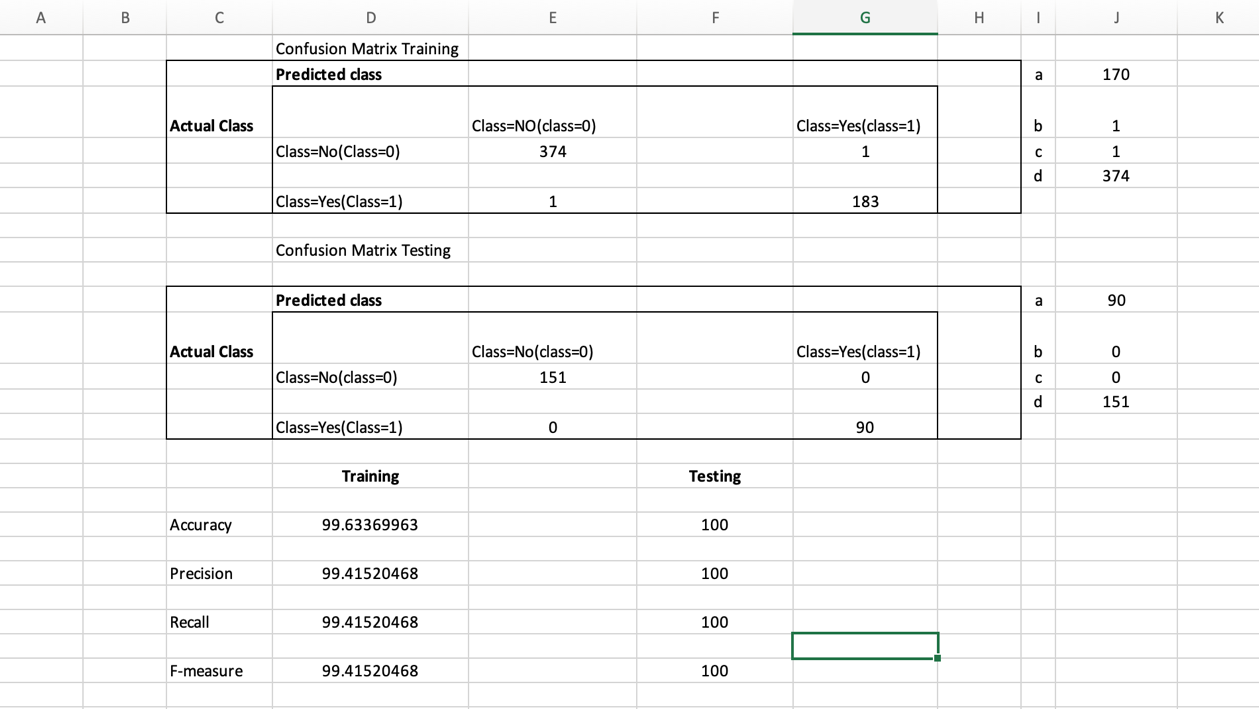
A diagram of a software system

Description automatically generated

**a.**

A screenshot of a computer

Description automatically generated



**b.**

Based on the metrics, here's a detailed evaluation and comparison between the CART model and the logistic regression model:

**CART Model Performance:**

**Training Set:**

- Accuracy: 99.633%

- Precision: 99.415%

- Recall: 99.415%

- F-measure: 99.415%

**Testing Set:**

- Accuracy: 100%

- Precision: 100%

- Recall: 100%

- F-measure: 100%

**Logistic Regression Model Performance:**

**Training Set:**

- Accuracy: 96.958%

- Precision: 98.265%

- Recall: 92.391%

- F-measure: 95.238%

**Testing Set:**

- Accuracy: 96.680%

- Precision: 97.674%

- Recall: 93.333%

- F-measure: 95.454%

**Comparison:**

**Accuracy:** The CART model outperforms the logistic regression model on both the training and testing sets, achieving perfect accuracy on the testing set.

**Precision:** Precision is also higher in the CART model, indicating that when it predicts a loan will default, it is more likely to be correct.

**Recall:** The recall is significantly higher for the CART model, which means it is better at identifying all actual default cases.

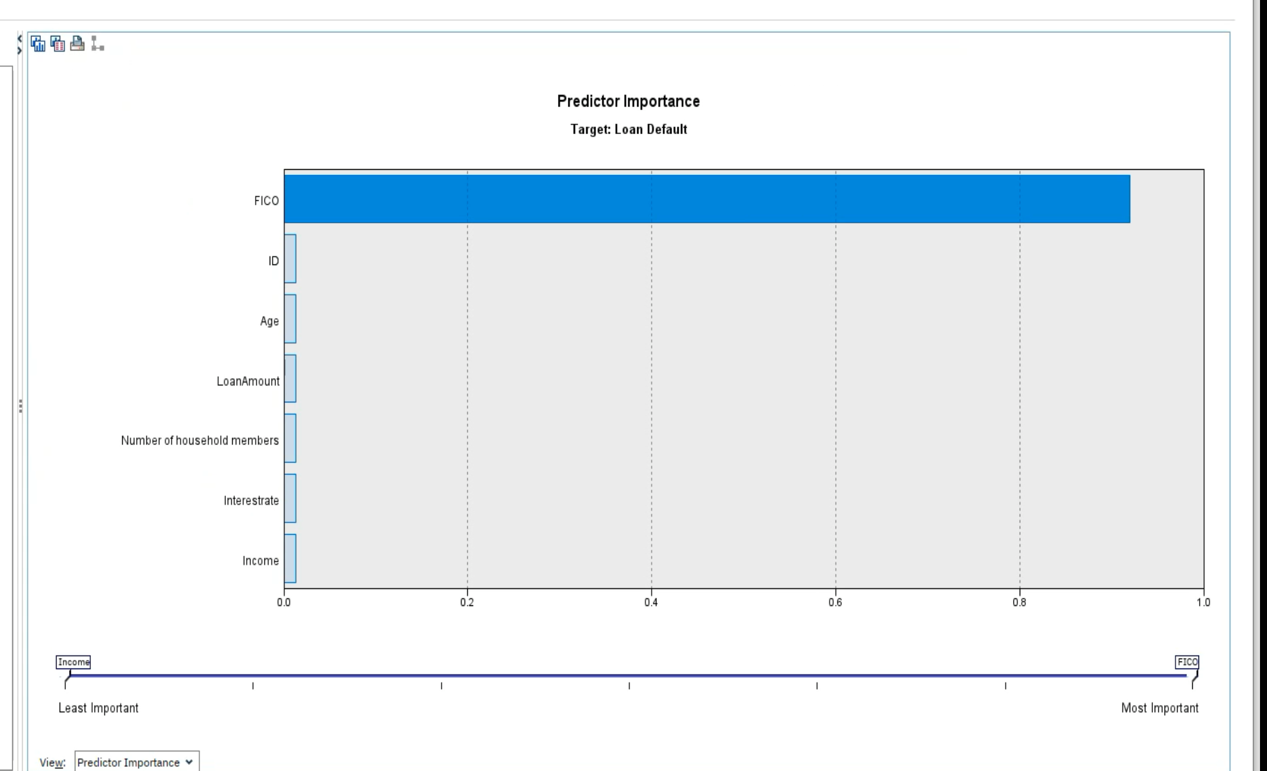
**F-measure:** The F-measure, which balances precision and recall, is higher for the CART model, showing a better overall harmonic balance between precision and recall.

**Evaluation:**

The CART model shows exceptionally high performance on both the training and testing sets. The perfect scores on the testing set suggest that it has learned to generalize extremely well to unseen data.

However, the perfect scores on the testing set, especially when contrasted with the already high scores on the training set, might indicate overfitting. Although overfitting typically results in lower test performance, it can also happen that the test set is not challenging enough or doesn't represent real-world complexity. The logistic regression model shows strong performance but not as high as the CART model.

**c.**



Gender does not appear as a predictor in the importance chart, it suggests that gender is not a significant factor in the decision-making process of the decision tree model when predicting loan default. This would imply that the model did not find a strong enough relationship between gender and the likelihood of defaulting to use it as a criterion for splitting the data.

Here's what this means in the context of discrimination:

**1. Lack of Discrimination by Gender:** The absence of gender as a predictor indicates that the model does not discriminate based on gender when making predictions. This is a positive outcome in terms of fairness and avoids ethical and legal issues that could arise from gender-based discrimination.

**2. Model Fairness:** The model appears to be fairer with respect to gender as it relies on other factors that are more strongly related to the likelihood of loan default. This suggests the decision tree is basing its predictions on financial behaviors and characteristics rather than on demographic attributes that could introduce bias.

**3. Model Selection: T**he fact that gender is not used as a splitting criterion does not affect the model's performance in terms of accuracy, precision, recall, and F-measure. However, it could positively influence the decision to use this model since it avoids potential biases that could be present if gender was an important predictor.

**4. Regulatory Compliance:** By not including gender, the model is likely to be more compliant with anti-discrimination laws and regulations that prohibit the use of certain characteristics in decision-making processes.

**5. Model Deployment:** If you are planning to deploy this model, the absence of gender as a predictor can make it easier to justify its use and can enhance the trust stakeholders have in the model's predictions.

**6. Further Analysis:** Even if gender is not a predictor, we still can evaluate the model's predictions across different gender groups to ensure there are no hidden biases. This can be done by analyzing the distribution of predictions and outcomes by gender post hoc.

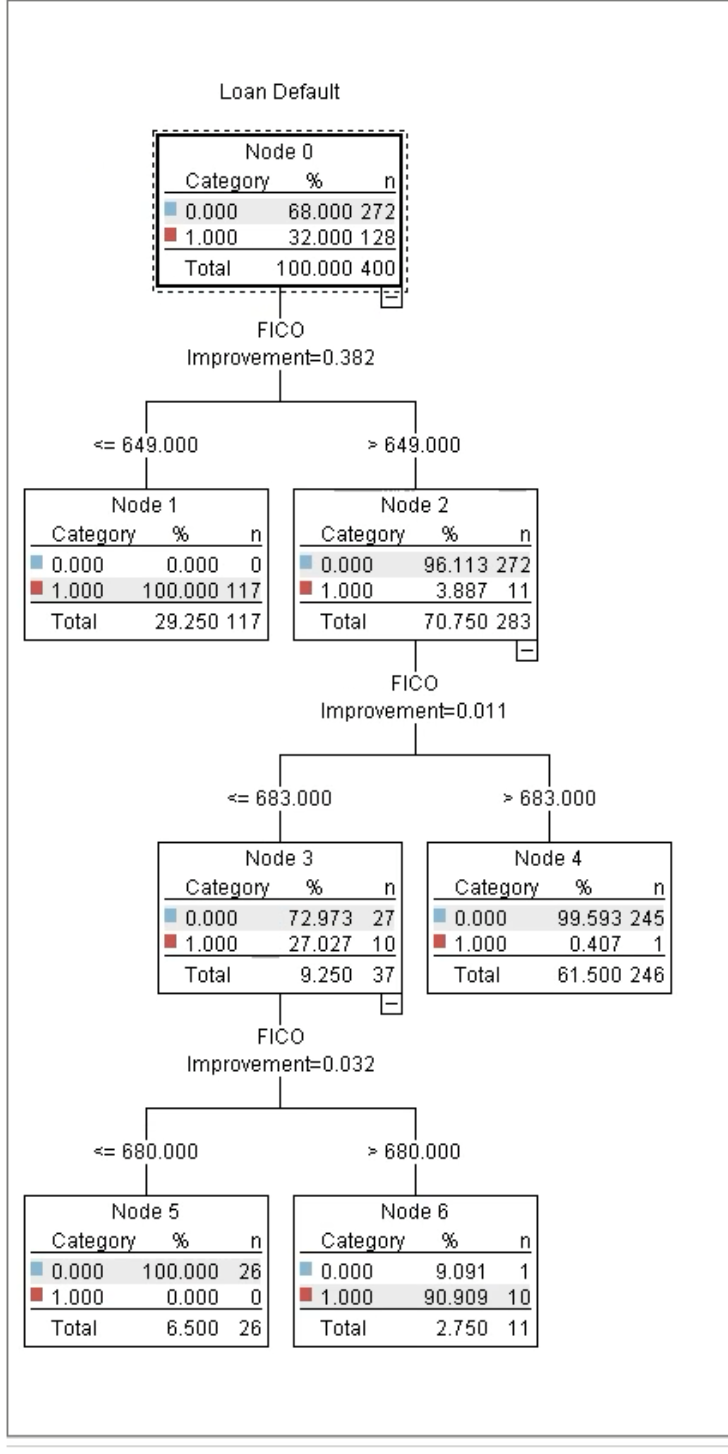
In summary, the lack of gender as a predictor in the decision tree is a positive indication that the model is not using this sensitive attribute to make predictions about loan default, which is good practice from both ethical and regulatory perspectives.

**d.**

**We have filled the table according to the decision tree below.**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Customer No | Number of household members | Income | Age | FICO | Gender | Loan Amount | Homeowner | Default or not |
| 1 | 1 | 75000 | 35 | 700 | 0 | 4500 | 0 | 0 |
| 2 | 1 | 75000 | 37 | 700 | 1 | 5500 | 0 | 0 |
| 3 | 1 | 45000 | 43 | 620 | 0 | 3500 | 0 | 1 |
| 4 | 1 | 45000 | 29 | 670 | 1 | 4000 | 0 | 1 |

To predict whether each of the new customers would default on their loans using the decision tree, we'll follow the rules inferred from the decision tree structure.



1. Check the FICO score of the customer.

2. Follow the path in the decision tree that corresponds to the customer's FICO score.

3. At each node, check if the customer's attributes match the condition.

4. Continue following the path until a terminal node (leaf), which will indicate the prediction of default or not.

For Customer 1 and Customer 2 with FICO scores of 700:

* The FICO score of 700 is greater than the threshold value (649) at the first node in decision tree.
* Hence, following the decision tree path, the next threshold value is 683. And the FICO score of customers 1 and 2 is greater than this.
* Since both customers have a FICO score above 649 and above the next node threshold of 683, they would likely be predicted not to default if the tree continues to use FICO as the primary splitter and if higher FICO scores are associated with a lower probability of default in your tree.

For Customer 3 with a FICO score of 620 and Customer 4 with a FICO score of 670:

* Customer 3 has a FICO score less than the first node threshold of 649, so following the branch for scores less than 649. Hence, customer 3 is likely to be default.
* Customer 4 has a FICO score that falls between the first threshold (649) and the second threshold (683). Hence following the branches, the FICO is less than 680. Hence, customer 4 is also default.

**References:**

* *Data Analysis, Statistical & Process Improvement Tools | Minitab*. (n.d.). Www.minitab.com. <http://www.minitab.com>
* OpenAI. (2024). *ChatGPT*. Chat.openai.com; OpenAI. <https://chat.openai.com>
* *Decision Tree Nodes*. (n.d.). Www.ibm.com. Retrieved March 12, 2024, from https://www.ibm.com/docs/en/spss-modeler/18.4.0?topic=trees-decision-tree-nodes