***“Predicting Personal Loan Approvals with Logistic Regression”***

***Business Analytics Using SPSS***

1. ***Bank loan approval data:***

**Source: https://www.kaggle.com/**

**File contains 14 columns and 5000 rows**

***Variables:***

* **ID: Customer ID**
* **Age: Customer Age**
* **Experience: Customer Experience**
* **Income: Income of the Customer**
* **Zip Code: Customer's residence zip code**
* **Family: No of Family members of the customer**
* **CC Avg: Credit Card Average Score**
* **Education: Education of the customer**
* **Mortgage: Mortgage taken or not taken by the customer**
* **Personal Loan: 0 = No personal loan given, 1 = personal loan given**
* **Securities Account: Having or not having a Securities Account**
* **CD Account: Having or not having a CD Account**
* **Online: Having or not having online banking**
* **Credit Card: Having or not having a credit card**

1. ***Objective:***

**The objective is to predict personal loan approval for customers using logistic regression. This analysis aims to identify key factors influencing loan approval and improve decision-making efficiency for the bank.**

1. ***Logistic Regression:***

**Logistic regression is a statistical method used to predict a binary outcome (e.g., yes/no, 0/1) based on one or more independent variables. In this case, the goal is to predict whether a customer will be approved for a personal loan (1) or not (0), making logistic regression an ideal choice.**

1. ***Assumptions for Logistic Regression:***

* **Binary Dependent Variable:** The dependent variable must be binary (e.g., 0/1, Yes/No).
* **Independence of Observations:** Each observation should be independent, with no duplicated data or correlations between observations.
* **No Multicollinearity:** Independent variables should not be highly correlated. Use a Variance Inflation Factor (VIF) or a correlation matrix.
* **Linearity of the Logit:** The relationship between the independent variables and the log-odds of the dependent variable should be linear.
* **Sufficient Sample Size:** Logistic regression requires an adequate sample size to ensure stable coefficient estimates.

1. ***Key Concepts:***

**Logit Function**

* The logit function is the natural logarithm of the odds of the probability PPP of the outcome:

**logit(P)=ln(P/(1−PP​))**

* Logistic regression models the relationship between the predictors (input variables) and the logit of the target variable.

**Sigmoid Function**

* Logistic regression uses the sigmoid function to map predicted values to probabilities between 0 and 1:

**σ(z)=1/ (1+e-z)**

**where z=β0+β1x1+β2x2+⋯+βnxn**

**Decision Boundary**

* A threshold (commonly 0.5) is applied to the predicted probability to classify the output:

y​={1 if P(y=1∣X)≥0.5 , 0 Otherwise

**Dependent Variable (Target)**

* **The binary outcome variable (yyy), which can take values 0 or 1.**

**Independent Variables (Features)**

* **The input variables (x1,x2,…,xnx\_1, x\_2, \dots, x\_nx1​,x2​,…,xn​) used to predict the target.**

**Coefficients (β\betaβ)**

* **Parameters of the model learned during training. They represent the relationship between features and the log-odds of the target.**

**Odds**

* **The ratio of the probability of success (P) to failure (1−P):**

**Odds=P/ (1−P)​**

**Log-Likelihood**

* **The logistic regression model maximizes the log-likelihood, which measures how well the model fits the data.**

1. ***Dependent and Independent Variables***
2. **Dependent Variable (Y):**

Personal Loan: (0 = Not Approved, 1 = Approved)

1. **Independent Variables (X):**
   * Age
   * Experience
   * Income
   * ZIP.Code
   * Family
   * Education(Categorial)
   * Securities.Account(Categorial)
   * CC Avg
   * Mortgage (Categorial)
   * CD Account (Categorial)
   * Online (Categorial)
   * Credit Card (Categorial)
2. ***Hypothesis:***

**Null Hypothesis (H0​):**

The independent variables (e.g., Age, Income, CC Avg, Education, etc.) have no significant impact on the likelihood of personal loan approval.

**H0:β1=β2=β3=......=βn=0**

This means the model does not improve prediction compared to a baseline.

**Alternative Hypothesis (H1​):**

At least one independent variable significantly impacts the likelihood of personal loan approval.

**H1: βi≠0, for at least one i**

This implies that the logistic regression model provides meaningful insights into the factors affecting loan approval and significantly improves prediction accuracy.

1. ***Logistic Regression Using SPSS***

**Steps:**

**--Analyse > Regression > Binary Logistic**

**--Dependent variable Box: Add Dependent Variable (Categorical)**

**--Block 1 of 1: Add Independent variables**

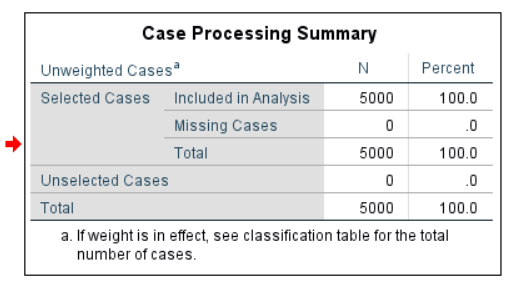
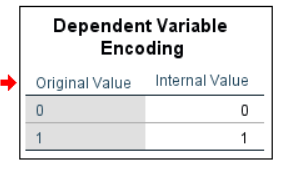
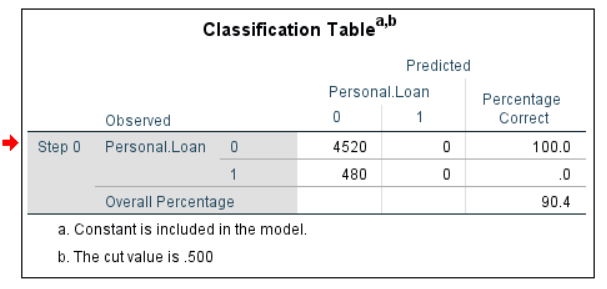
**--Method: Enter**

**--Categorial >Add those variables which are categorial.**

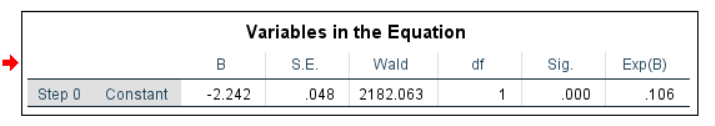
**--Save > Click On probability and Group membership**

**--Options >Hosmer- Lemeshow Goodness-of-fit**

**+ Include Constant in model.**

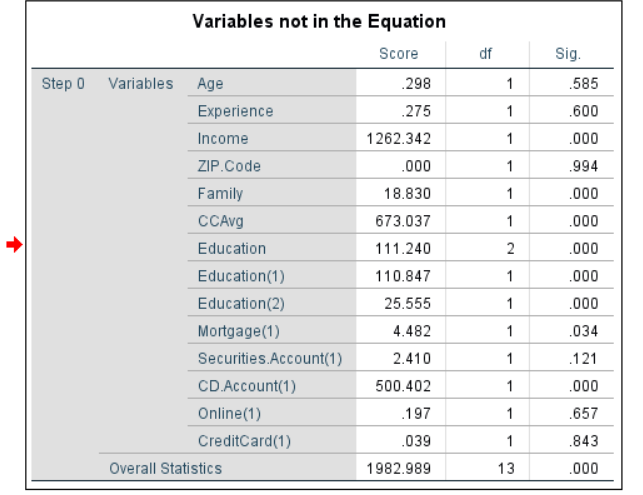
1.  *Result and Interpretation:*
2. ***Classification Table***

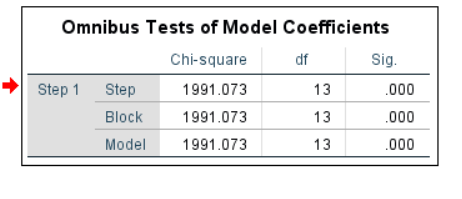
* Result = 0: All 4520 cases were correctly predicted (100% accuracy).
* Result = 1: None of the 480 cases were correctly predicted (0% accuracy).
* Overall accuracy = 90.4%, based on predicting all cases as the majority class (Result = 0).

1. ***Variables in Equation:***

**Constant (Intercept):**

* + **B = -2.242**, significant (p < 0.05), with an odds ratio **(Exp(B)) of 0.106**.
  + This indicates that without any predictors, the odds of success (Result = 1) are **10.6%** of the odds of failure (Result = 0).

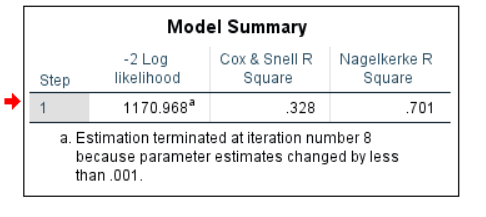
1. ***Variables not in the Equation:***
   * Predictors (Income, Family, Education(cat.), CD Account(cat.), CC Avg) have significant scores (p < 0.05), indicating they are likely to improve the model.
   * Predictor (Age, Experience, ZIP. Code, Family (3), Mortgage (1), Securities.Account(1), Online (1), Credit Card (1)) has significant scores (p >0.05), indicating it will not fit well into the model equation.
2. ***Omnibus Test of Model Coefficient:***



**Chi-square = 1991.073**

p < 0.001: The model with predictors (Income, Experience, Education(cat.), ZIP.Code, Family(cat.), Securitirs.Account, Mortgage(cat.), CD Account(cat.),CC Avg, Online, CreditCard, Age) significantly improves over the null model.

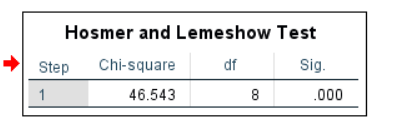
1. ***Model Summary:***



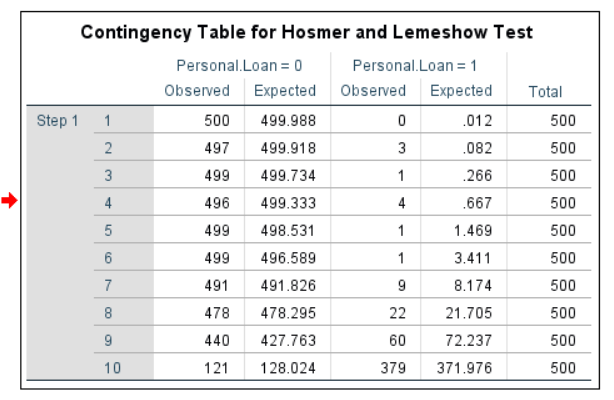
* **-2 Log Likelihood = 1170.968**: Indicates the goodness-of-fit; lower values signify a better fit.
* **Cox & Snell R² = 0.328, Nagelkerke R² = 0.701:**

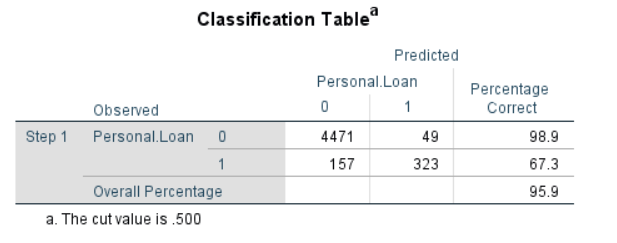
model explains 32.8-70.1% of the variance in the dependent variable.

1. ***Hosmer And Lemeshow Test:***

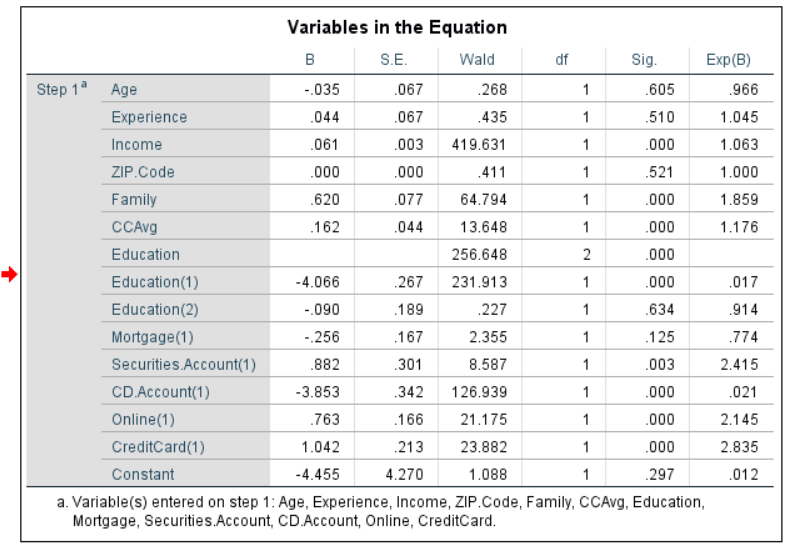
**Chi-Square (46.543):** This statistic tests whether the independent variables collectively improve the model compared to the null model.

**p-value (0.000):** A p-value < 0.05 indicates that the improvement in model fit is statistically significant.



1. ***Step 1: Classification Table:***

* Result = 0: 4471 out of 4520 cases were correctly predicted (98.9% accuracy).
* Result = 1: 157 out of 480 cases were correctly predicted (667.3% accuracy).
* Overall accuracy = 95.9%, a significant improvement from Step 0 (90.4%).

1. ***Variables in the Equation:***

**Age:**

* B = -.035, Exp(B) = 0.966: The odds of success (Result = 1) Decrease by a factor of 0.966 (96.6% Decrease).
* Not significant (0.605 > 0.001).

**Experience:**

* B =.044, Exp(B) = 1.045: For increase in income, the odds of success increase by a factor of 1.045 (104.5% increase).
* Not significant (0.510 < 0.001).

**Income:**

* B =0.061, Exp(B) = 1.063: For increase in income, the odds of success increase by a factor of 1.063 (106.3% increase).
* Highly significant (0.000 < 0.001).

**ZIP.Code:**

* B =.000, Exp(B) = 1.000: For Different location, the odds of success increase by a factor of 1.00 (100% increase).
* Not significant (0.521 < 0.001).

**Family:**

* B =0.620, Exp(B) = 1.859, With Increase Number of Family members, the odds of success increase by a factor of 1.859 (185.9% increase).
* Highly significant (0.000 < 0.001).

**Education (1):**

* B = -4.066, Exp(B) = .017: The odds of success (Result = 1) Decreases by a factor of 0.017 (1.7%).
* Highly significant (p < 0.001).

**Education (2):**

* B = -0.090, Exp(B) = .914: The odds of success (Result = 1) Decreases by a factor of .914 (91.4%).
* Not significant (0.634 >0.001).

**Mortgage (1):**

* B = -0.256, Exp(B) = 0.774: One who have to pay mortgage, the odds of success Decreases by a factor of 0.744 (74.4%).
* Not significant (0.125 > 0.001).

**Securities.Account (1):**

* B = 0.882, Exp(B) = 2.415: The odds of success (Result = 1) Decreases by a factor of 2.415 (241.5%).
* Not significant (0.003 >0.001).

**CD Account (1):**

* B = -3.853, Exp(B) = .021: The odds of success (Result = 1) Decreases by a factor of 0.021 (2.1%).
* Highly significant (0.000 < 0.001).

**CC Avg:**

* B = .162, Exp(B) = 1.176: The odds of success (Result = 1) increase by a factor of 1.176 (117.6% increase).
* Highly significant (0.000 >0.001).

**Online (1):**

* B = .763, Exp(B) = 2.145: The odds of success (Result = 1) increase by a factor of 2.145 (214.5% increase).
* Highly significant (0.000 >0.001).

**Credit Card (1):**

* B = 1.042, Exp(B) = 2.835: The odds of success (Result = 1) increase by a factor of 2.835 (283.5% increase).
* Highly significant (0.000 >0.001).

**Constant:**

* B = -4.455, Exp(B) = 0.012, The log-odds of success are low when both predictors are 0.

1. ***Final Interpretation:***

**The equation becomes:**

logit(P)=−4.455 −0.035(Age) +0.044(Experience) +0.061(Income) +0.000(ZIP Code) +0.620(Family) +0.162(CCAvg) −4.066(Education [1]) −0.090(Education [2]) −0.256(Mortgage [1]) +0.882(Securities.Account [1]) −3.853(CD.Account[1]) +0.763(Online [1]) +1.042(CreditCard [1])

* Variables with significant p-values (Sig. < 0.05) are more likely to contribute meaningfully to the prediction of Personal loan . These include Income, Family, CCAvg, Education [1], Securities.Account[1], CD.Account[1], Online[1], and CreditCard[1].
* Exp (B) values give the odds ratio for each predictor, showing how the odds change with a one-unit increase in the predictor.

**1.Significance:**

* Both predictors (Income, Family, Education(1),CD Account(1), CC Avg, Online(1), Credit Card(1)) significantly impact the odds of success (p < 0.001).
* The model explains a large portion of the variance (Nagelkerke R² = 0.701).

**2.Effect Sizes:**

* Income, Family, CC Avg, Online Banking and having Credit Card, have stronger effect on the odds of success (Exp(B)> 1) respectively compared to Other significant Independent Variables.

**3.Model Performance:**

* The overall predictive accuracy improved from 90.4% (null model) to 95.9% (Step 1).
* The Hosmer and Lemeshow test confirm a good model fit.
* Income, Family, CC Avg, Online Banking and having Credit Card are strong predictors while Education (1), CD Account (1), are Moderate but significant predictor of success in the dependent variable.
* The logistic regression model performs well, with high explanatory power and predictive accuracy.