### Research with solid fill5.3 Hands-on Case Study:

Logical Regression Modelling

*Download Source Data Set from GitHub link* : <https://github.com/Smartbrain2024/Mastering_AI_2.git>

**Chapters/Chp\_05/5.3/Hands\_on/AutoLend.csv**

Use Linear regression & Logistic classification to build models to be used for redefining loan approval and loan pricing strategy to minimize loss for an American auto finance company.

**AutoFinc** is a US based provider of auto financial services including banking, vehicle finance & vehicle insurance for cars, bikes and commercial vehicles servicing retail as well as corporate customers.

Auto financers aim to have **high approval rates** to increase overall efficiency and strengthen relationship with auto dealerships. At the same time, they try to **mitigate losses and defaults**. This overall is a tricky balance to achieve.

**Auto FinTech is a market leader in the car financing in the retail customer segment with a 22% market share, but the vertical has seen 17% more default rate and 12% more loss due to default as compared to what was predicted.**

**The Risk Analytics team is assigned the task to come out with new decision rules for loan approval and loan pricing to minimize loss.**

**Problem Overview :**

* To produce new decision rules for loan approval and loan pricing, the team will need to estimate the expected loss for each loan applicant using the below formula. This will depend on factors like an applicant’s demographic, socio-economic profiles, and application details (credit amount, duration etc.).
* Expected Loss = PD x EAD
* PD = the probability that a loan applicant will default
* EAD (Exposure at Default) = the outstanding loan amount at the time of default
* For this, the Risk Analytics team needs to:
* Build a logistic regression model which will predict probability of default (PD)
* Build a linear regression model to predict the exposure at default (EAD)

**Data Dictionary:**

The dataset contains demographic, socio-economic, credit amount and loan duration for 1000 customers.

The data contains whether a customer defaulted and if he/she defaulted what the outstanding amount was.

The data has been checked for missing values, duplicate rows, and data type. We can move directly to EDA step.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Age | Numerical: Age in years |
| Sex | Categorical: male, female |
| Job | Categorical: unskilled and non-resident, unskilled and resident, skilled, highly skilled |
| Housing | Categorical: own, rent, or free |
| Saving accounts | Categorical: little, moderate, quite rich, rich |
| Checking account | Categorical: little, moderate, rich |
| Duration | Numerical: Duration for which the credit is given in months |
| Credit Amount | Numerical: Amount of loan credit in USD |
| Risk | Categorical: 0 – Person is not at risk, 1 – Person is at risk(defaulter) |

A graph of a person

Description automatically generated with medium confidence

A diagram of a graph

Description automatically generated

The distribution of “Age” is right skewed. Youngest 25% of customers are 19-27 years old; 50% are under 33 years. 33 years is median age and 35 years mean.

The distribution of “Duration” is also right skewed. 75% of the loans have duration 4-24 months. 18 months is the median duration and 21 months mean.

A graph of blue bars

Description automatically generatedA diagram of numbers and a line

Description automatically generatedA black background with white text

Description automatically generated

A graph of a graph

Description automatically generated

The distribution of “Exposure at Default” is normally distributed with median at 12806 USD and mean at 12538 USD.

The distribution of “Credit Amount” is normally distributed with 17949 USD median and 17881 USD mean.

The distribution of “Credit Amount” is normally distributed with 17949 USD median and 17881 USD mean.

A graph of different colored squares

Description automatically generated with medium confidenceEDA – Bivariate Analysis

A graph of different colored rectangles

Description automatically generatedA blue and orange bar graph

Description automatically generated

Second and third quartiles of duration of defaulters is much more than that of non-defaulters. This shows that customers with high duration are more likely to default.

* “Age” distribution is similar for defaulters and others. Defaulters are slightly younger with a median at 31 years.
* There is not a lot of difference in defaults depending on credit amount.

A blue and orange bar graph

Description automatically generated

* Females see a higher percentage of defaulters at 30% as compared to 21% for males.
* Default % decreases as skill level increases.
* Default % goes higher as the saving account health declines.
* Default % goes higher as the checking account health declines.

##### Model Building – Logistic Regression

A diagram of a logistic regression model

Description automatically generated

* The random state can be set to 1 while using the ‘Partitioning’ node.
* Target Variable is “Risk”

*Model Evaluation – Accuracy Statistics*

* We want to predict whether a customer will default (1) on his/ her loan payment using the above information.
* This model can make wrong predictions as:
  + It predicts a non-defaulter (0) as a defaulter (1) – **False Positive**
  + It predicts a defaulter (1) as a non-defaulter (0) – **False Negative**
* Auto Fintech would want the model to reduce its False Negatives as a defaulter has an excessive cost to the company. So, a good model should have a **high Recall.** While, more False Positives would result in higher loan rejections (which is loss of business)

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

A screenshot of a table with numbers and text

Description automatically generated

* Our model has given a high overall accuracy of 88% on test data.
* We are getting a Recall of 64.5% and an F1 Score of 69% for Test data. F1 lets us minimize both False Positive & False Negative.
* The Area under ROC curve is 0.878 for Test data.

A graph with a line

Description automatically generated

A screenshot of a table

Description automatically generated

* P-value of a variable indicates if the variable is significant or not. If we consider the significance level to be 0.05 (5%), then any variable with a p-value less than 0.05 would be considered significant.

###### Insights & Recommendations:

* From our logistic regression model, we identified that Job, housing & savings & checking accounts health is a significant predictor of a customer being a defaulter.
  + Customers whose profile gives a high-risk probability should not be given high duration loans.
  + Customers with skilled and unskilled/non-resident jobs are elevated risk customers.
  + Customers owning a house are less likely to default while those with free or rented housing are at elevated risk of default. The bank should have thorough KYC documentation (Know Your Customer) to keep more details about such customers like hometown addresses, etc. to be able to track them as not owning a home is substantial risk.
  + Customers with little amount in saving accounts are more likely to default as compared to rest.
  + Customers with little amount in checking accounts are most likely to default as compared to moderate and rich.
* We saw in our analysis that customers with a little amount. The bank can be stricter with its rules or interest rates to compensate for the risk.
* Auto Fintech should try to sell more shorter duration loans as they have lesser chance of default.
* Moderate savings amount customers who default tend to have a high Exposure at Default.
* Exposure at default is proportional to credit amount and hence policy can have an upper limit on credit amount for high & moderate risk customers.