**Credit Risk Classification Analysis.**

**Credit** is “created” when one party receives resources from another party, but payment is not expected until some contracted date (or dates) in the future. Credit analysis is a process undertaken by lenders to **understand the creditworthiness of a prospective borrower**, meaning how capable (and how likely) they are of **repaying principal and interest obligations.** The most popular credit analysis framework for risk rating is the 5C’s model which involves; - **Character, Capacity, Capital, Collateral,** and **Conditions**. The borrower, also known as the debtor, could be an individual or a business entity; the former is referred to as retail (or personal) lending, and the latter is what’s known as **commercial lending**. Lenders, also known as creditors, employ a variety of **qualitative** and **quantitative techniques** (including **financial risk models**) when conducting credit analysis in order to quantify and effectively evaluate the risk's status.

Credit professionals analyzing a prospective borrower will employ a variety of **qualitative** and **quantitative techniques**.

Qualitative techniques include trying to understand risks in the **external environment**, like the current and projected **micro** and **macro-economic factors.** Mostly **PESTEL** (political, economic, social, technological, environmental, and legal) factors which are analyzed and screened to ascertain the credit risks expected. For **commercial lenders**, specifically, they’ll also want to understand business characteristics – like the borrower’s competitive advantage(s) and industry trends (using frameworks like **SWOT** **Analysis** - (strengths, weakness, opportunities, and threats) and **Porter’s 5 Forces** which includes **competitive rivalry, supplier power, buyer power, threat of substitutions** and **threat of entry**), respectively. Entity portfolio management experience is another very important consideration.

Quantitative elements of the analysis include assessing financial ratios using risk models, understanding financial projections, employing sensitivity analysis, and evaluating the strength of any physical collateral that could serve as security against credit exposure.

To Analyze and predict the credit risks of the borrowers, I used a **supervised algorithm** known as the **Logistics Regression Model** where data was split into training and testing sets. The primary data available and used is the borrower's loan size, interest rates, borrower's income, debt-to-income ratio, number of accounts, derogatory marks, and total debts.

In the datasets, the **loan status** is healthy when the value is equivalent to zero while one means the loan has a high risk. So, **y.value \_count( )** was used to determine the number of healthy loans denoted as “0” and unhealthy loans denoted as “1”. As determined in the dataset 75,036 loans are healthy with fewer/less risks compared with 2500 loans which are high-risk loans.

After importing the lending data containing features (independent variables) and a binary outcome variable (dependent variable). The data was cleaned, reprocessed, and then the train-test model was used to split the data. The Logistics Regression Model as a classification algorithm was used to analyze the data. A confusion matrix was also employed to evaluate the performance of the risk analysis. Using the confusion matrix a prediction was made compared with the actual labels on the ground.

Using confusion matrix various performance metrics are derived, such as accuracy, precision, recall (sensitivity), specificity, F1-score, etc., which helps in understanding how well the model is performing and where it may need improvement. It breaks down the model predictions into the following categories; -

1. True Positives (TP) – cases where the positive class predicted correctly actual label is positive.
2. True Negative (TN) – cases where the negative class predicted correctly actual label is negative.
3. False Positive (FP) – cases where the positive class is predicted incorrectly (false alarm) while the actual label is negative.
4. False Negative (FN) – cases where the negative class is predicted incorrectly(miss) while the actual label is positive.

So, what does the Confusion Matrix tell us?

The **True Positives** and **True Negatives** indicate accurate predictions.

The **False Positives** indicate the model incorrectly predicted the positive class.

The **False Negatives** indicate the model failed to identify and predict the positive class.

The following are the most important terms used in the confusion matrix model; -

1. **Accuracy**- this measures the total number of correct classifications divided by the total number of cases. Accuracy = TP+TN / TP+TN+FP+FN
2. **Recall/Sensitivity** - this measures the total number of true positives divided by the total number of actual positives. Recall = TP / TP+FN or TP /Actual Results.
3. **Precision**- this measures the total number of true positives divided by the total number of predicted positives. Precision = TP / TP+FP or TP/Predictive Results
4. **Specificity** - this measures the total number of true negatives divided by the total number of actual negatives. Specificity = TN / TN+FP
5. **F1 Score** - is a single metric that is a harmonic mean of precision and recall. F-Score = 2\*Recall\*Precision/Recall+ Precision

The following are the results predicted; -

**Precision:** Precision is the ratio of true positive predictions to the total number of positive predictions made by the model. It measures the accuracy of positive predictions. In this case, for the "healthy loan" class, the precision is 1.00, meaning all positive predictions for "healthy loan" were correct. For the "high-risk loan" class, the precision is 0.87, indicating that 87% of the positive predictions for "high-risk loan" were correct.

**Recall/Sensitivity**: Recall is the ratio of true positive predictions to the total number of actual positive instances in the dataset. It measures the ability of the model to identify all relevant instances. A recall of 1.00 means the model didn't miss any positive instances. In this case, for the "healthy loan" class, the recall is 1.00, indicating that all instances of "healthy loan" were correctly identified. For the "high-risk loan" class, the recall is 0.89, meaning 89% of the actual "high-risk loan" instances were identified by the model.

**F1-Score:** The F1-score is the harmonic mean of precision and recall. It provides a balance between precision and recall, giving more weight to low values. It's a single metric that combines both precision and recall into one. In this case, the F1-score for a "healthy loan" is 1.00, and for a "high-risk loan" is 0.88.

**Support:** Support refers to the number of actual occurrences of each class in the dataset.

**Accuracy**: Accuracy is the ratio of correct predictions to the total number of predictions made by the model. It measures the overall correctness of the model. In this case, the overall accuracy is 0.99, meaning the model correctly predicted 99% of the instance.

The following are some of the machine learning models used; -

**Supervised Learning (Task-driven):**

* **Classification**; - Naïve Bayes, Nearest Neighbor, Discriminant Analysis.
* **Regression;** -Linear Regression, Generalized Additive Model, Gaussian Process,
* **Classification or Regression**; - Support vector Machines, Decision Trees, Ensemble Methods/Gradient Boosting Machines, and Neural Networks.

**Unsupervised Learning (Data-driven):**

* **Clustering;** - K-Means, K-Medoids, Fuzzy C-Means, Hierarchical, Gaussian Mixture, Hidden Markov Model

**Reinforced Machine Learning** (Learn from errors).

In my opinion, for credit risk analysis I would recommend Logistics Regression Analysis because it is simple, efficient, provides probabilities, and is very effective on binary classifications. Decision Trees can be difficult to interpret and also not suitable for linear relationships.

First and foremost, it is very important to predict both 1’s and 0’s, because by predicting the 1’s its easier to identify the 0’s and vice versa.