# Module 20 Challenge

### **Background**

In this Challenge, you’ll use various techniques to train and evaluate a model based on loan risk. You’ll use a dataset of historical lending activity from a peer-to-peer lending services company to build a model that can identify the creditworthiness of borrowers.

### **Before You Begin**

1. Create a new repository for this project called credit-risk-classification. **Do not add this homework to an existing repository.**
2. Clone the new repository to your computer.
3. Inside your credit-risk-classification repository, create a folder titled "Credit\_Risk."
4. Inside the "Credit\_Risk" folder, add the credit\_risk\_classification.ipynb and lending\_data.csv files found in the "Starter\_Code.zip" file.
5. Push your changes to GitHub.

### **Files**

Download the following files to help you get started:

* [Module 20 Challenge files](https://static.bc-edx.com/data/dl-1-2/m20/lms/starter/Starter_Code.zip)
* [Links to an external site.](https://static.bc-edx.com/data/dl-1-2/m20/lms/starter/Starter_Code.zip)

### **Instructions**

The instructions for this Challenge are divided into the following subsections:

* Split the Data into Training and Testing Sets
* Create a Logistic Regression Model with the Original Data
* Write a Credit Risk Analysis Report

#### **Split the Data into Training and Testing Sets**

Open the starter code notebook and use it to complete the following steps:

1. Read the lending\_data.csv data from the Resources folder into a Pandas DataFrame.
2. Create the labels set (y) from the “loan\_status” column, and then create the features (X) DataFrame from the remaining columns.  
   **NOTE**A value of 0 in the “loan\_status” column means that the loan is healthy. A value of 1 means that the loan has a high risk of defaulting.
3. Split the data into training and testing datasets by using train\_test\_split.

#### **Create a Logistic Regression Model with the Original Data**

Use your knowledge of logistic regression to complete the following steps:

1. Fit a logistic regression model by using the training data (X\_train and y\_train).
2. Save the predictions for the testing data labels by using the testing feature data (X\_test) and the fitted model.
3. Evaluate the model’s performance by doing the following:
   * Generate a confusion matrix.
   * Print the classification report.
4. Answer the following question: How well does the logistic regression model predict both the 0 (healthy loan) and 1 (high-risk loan) labels?

#### **Write a Credit Risk Analysis Report:** Write a brief report that includes a summary and analysis of the performance of the machine learning models that you used in this homework. You should write this report as the README.md file included in your GitHub repository.

Structure your report by using the report template that Starter\_Code.zip includes, ensuring that it contains the following:

1. **An overview of the analysis:** Explain the purpose of this analysis.
2. **The results:** Using a bulleted list, describe the accuracy score, the precision score, and recall score of the machine learning model.
3. **A summary:** Summarize the results from the machine learning model. Include your justification for recommending the model for use by the company. If you don’t recommend the model, justify your reasoning.

### **Requirements**

#### **Split the Data into Training and Testing Sets (30 points)**

To receive all points, you must:

* Read the lending\_data.csv data from the Resources folder into a Pandas DataFrame. (5 points)
* Create the labels set (y) from the “loan\_status” column, and then create the features (X) DataFrame from the remaining columns. (10 points)
* Split the data into training and testing datasets by using train\_test\_split. (15 points)

#### **Create a Logistic Regression Model (30 points)**

To receive all points, you must:

* Fit a logistic regression model by using the training data (X\_train and y\_train). (10 points)
* Save the predictions on the testing data labels by using the testing feature data (X\_test) and the fitted model. (5 points)
* Evaluate the model’s performance by doing the following:
  + Generate a confusion matrix. (5 points)
  + Generate a classification report. (5 points)
  + Answer the following question: How well does the logistic regression model predict both the 0 (healthy loan) and 1 (high-risk loan) labels? (5 points)

#### **Write a Credit Risk Analysis Report (20 points)**

To receive all points, you must:

* Provide an overview that explains the purpose of this analysis. (5 points)
* Using a bulleted list, describe the accuracy, precision, and recall scores of the machine learning model. (5 points)
* Summarize the results from the machine learning model. Include your justification for recommending the model for use by the company. If you don’t recommend the model, justify your reasoning. (10 points)

#### **Coding Conventions and Formatting (10 points)**

To receive all points, you must:

* Place imports at the top of the file, just after any module comments and docstrings and before module globals and constants. (3 points)
* Name functions and variables with lowercase characters, with words separated by underscores. (2 points)
* Follow DRY (Don’t Repeat Yourself) principles, creating maintainable and reusable code. (3 points)
* Use concise logic and creative engineering where possible. (2 points)

#### **Code Comments (10 points)**

To receive all points, your code must:

* Be well commented with concise, relevant notes that other developers can understand. (10 points)