Module 20 Challenge

Start Assignment

Due Sep 7 by 11:59pm **Points** 100
Submitting a text entry box or a website url

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

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)

Grading

This project will be evaluated against the requirements and assigned a grade according to the following table:

Grade	Points
A (+/-)	90+
B (+/-)	80–89
C (+/-)	70–79
D (+/-)	60–69
F (+/-)	< 60

Submission

You are required to submit the URL of your GitHub repository for grading.

NOTE

> Projects are requirements for graduation. While you are allowed to miss up to two Challenge assignments and still earn your certificate, projects cannot be skipped.

IMPORTANT

> **It is your responsibility to include a note in the README section of your repo specifying code source and its location within your repo**. This applies if you have worked with a peer on an assignment, used code in which you did not author or create sourced from a forum such as Stack Overflow, or you received code outside curriculum content from support staff such as an Instructor, TA, Tutor, or Learning Assistant. This will provide visibility to grading staff of your circumstance in order to avoid flagging your work as plagiarized.
>
> If you are struggling with a Challenge or any aspect of the curriculum, please remember that there are student support services available for you:
>
> 1. Office hours facilitated by your TA(s)
> 2. Tutor sessions (sign up)
> 3. Ask the class Slack channel/get peer support
> 4. AskBCS Learning Assistants

References

Data for this dataset was generated by edX Boot Camps LLC, and is intended for educational purposes only.