1.1 Lesson Plan: Predicting Credit Defaults with Amazon SageMaker

Overview

Today's Class will train a Random Forest on data provided by Home Credit to determine if a borrower is likely to default on a loan. The exploratory data analysis and model training will be conducted on a Jupyter Notebook on Amazon SageMaker. The module files can be found at this github repository.

Learning Objectives

At the end of the session, learners will be able to:

- Upload dataSets to an Amazon S3 bucket
- Retrieve data from a S3 bucket into a Jupyter notebook on SageMaker
- Understand how an exploratory data analysis is beneficial to train a model
- Understand how imputation is used to estimate missing values
- Understand how to create a random forest model

Instructor Notes

Students will be expected to clone the data from the class repository and follow along to upload the files into their own S3 bucket. Once a notebook has been created proceed to introduce the topics on exploratory data analysis using the slides.

Demonstrate an example exploratory analysis using the mtcars dataset. Split students into breakout rooms where they will try an exploratory analysis on the creditDefaultTrain dataset. Upon returning from the breakout rooms have students share their findings. Proceed to finish the class by training the random forest model on creditDefaultTrain and then answer any questions.

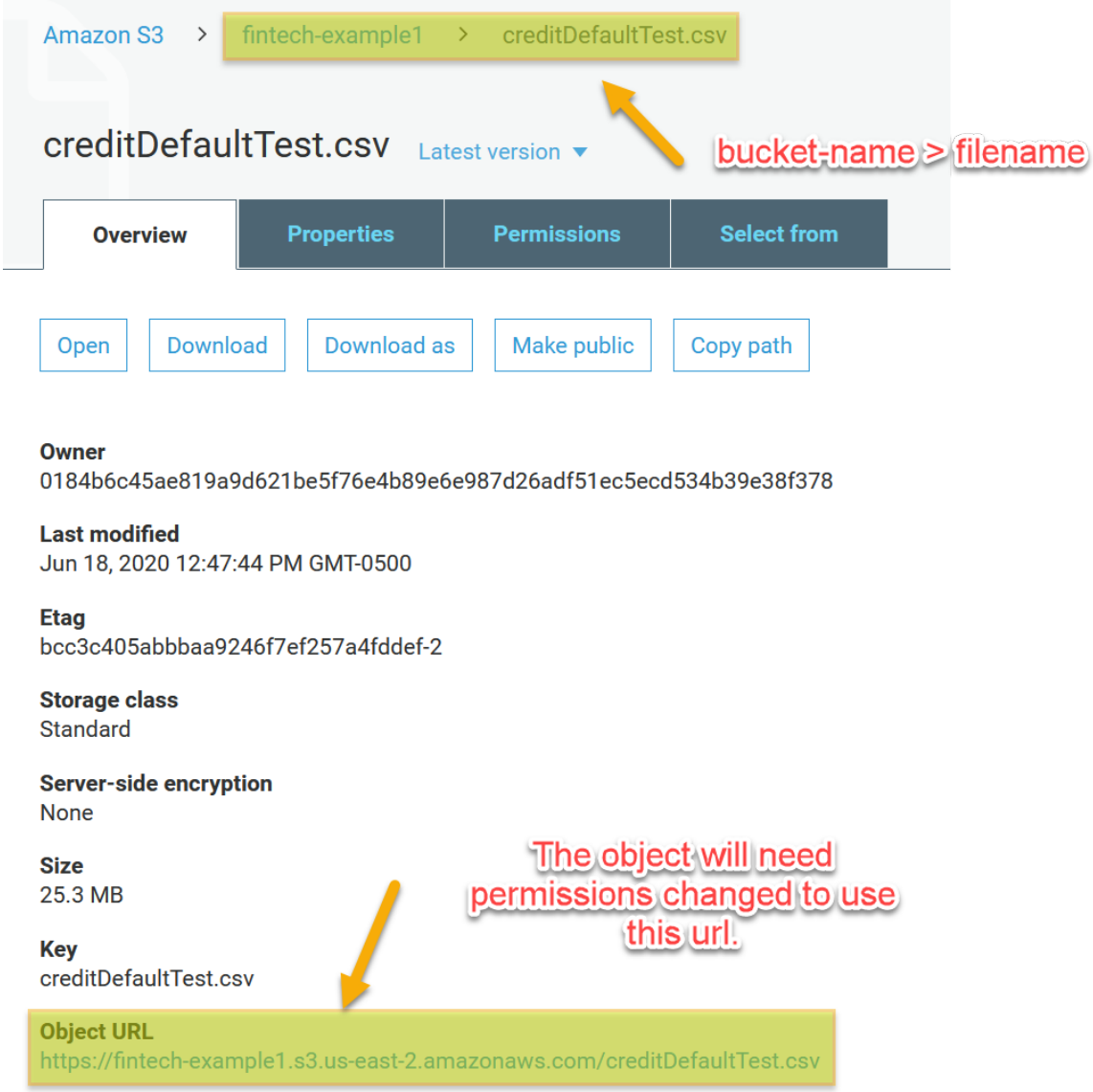
Module Student Dependencies

- AWS account
- github account
- python proficiency

Potential Student Environment Issues

- Students maybe confused in their scripts between using the object URL and using an s3:// url with the bucket name. If they try to use the object URL the file will not be accessible by SageMaker until the security settings are changed to make it public. Best practice would be to use the s3:// url and bucket name as the file does not have to be made public.

> **Tip**: If time permits and a student is having environment difficulties you may ask them to share their screen while you help them resolve it. Other students will benefit who are having the same issue.



s3:// urls are thus created:

```
s3URL = "s3://fintech-example1/creditDefaultTest.csv" #s3://bucket-name/fileName
```

Git LFS

If a repository needs to be configured for LFS after it has been created it maybe done so in the following way:

```
git lfs install
git lfs track "*.csv" #type of file to track
git add .gitattributes

git add .
git commit -m "Added lfs"
git push origin master
```

Additional Reading

- An Introduction to Statistical Learning - Examples are in R and the theory is explained concisely.

0. Class Do: Interview Question Warm-Up

(5 mins before class - first 2 mins of class)

Open the slideshow for today's class and begin the weekly presentation with the first slide.

> **This week's question:** What are some advantages and disadvantages for using a cloud service to train machine learning models?

Allow the question to be on the screen 5 mins prior to the start of class as students join the session. Ask the class to answer the question as they complete the pre-lecture temperature check.

Possible answers to this week's question:

Advantages

- Maybe cheaper than setting up hardware at the organization.
- Quicker to set up and use
- Easy to scale if more resources are needed

Disadvantages

- Businesses may not allow external services to access sensitive data
- New authentication methods could be difficult to integrate into an enterprise's existing user authentication system.

0. Instructor Do: Temperature Check

(5 mins before class - first 2 mins of class)

Using the Zoom Polling feature launch a poll for the class to identify where the class as a group is comfortable with the material. Do this while people are joining and during this time the TA may take attendance.

Poll Text:

Select all of the topics that you feel prepared to apply outside of the class from this week's lesson:

- Accessing data from an Amazon S3 bucket : **A**
- Creating notebooks in SageMaker : **A**
- Pearson's correlation coefficient : **B**
- Label Encoding : **B**
- One-hot Encoding : **B**
- Pandas dtypes : **B**
- Imputation : **C**
- Random Forest : **C**

A. S3 data upload and SageMaker notebook creation - Everyone (15 minutes)

B. Exploratory Data Analysis

C. Imputation and Model Training