Challenge Instructions

The challenges in this course are based on a real-world problem in which you must create a predictive machine learning model and enter it in a competition with your fellow students. Your first challenge is to explore the data and answer some basic questions about it. These questions will help you get started exploring the data, and your answers will be graded.

Your grade for this challenge accounts for 25% of your overall grade for the course.

The competition for this course is provided by DrivenData, an organization dedicated to using cutting-edge data science to address some of the world's biggest social challenges. To get started, you must:

1. Go to the competition site at www.datasciencecapstone.org and register using the same email address you used for your edX account.
2. Sign up for the competition at https://datasciencecapstone.org/t/data-science/latest/.
3. Review the information in the competition site.
4. Download the data, put it in a folder on your computer, and unzip it.
5. Explore the data in the ***Training set values (train_values.csv)*** and ***Training set labels (train_labels.csv)*** datasets.
6. Answer the questions in the next topic of this section.

You can explore the customer data using tools of your choice. Potential tools that you could use include:

- Microsoft Excel
- Microsoft Power BI
- R
- Python

- Microsoft Azure Machine Learning
- Spark

Analyze the Data

Based on your analysis of the original, unmodified training data:

Answer the following questions based on summary statistics you have calculated from the training dataset. **Be sure to round your answers to ONE *digit after the decimal point*** (e.g., **5.5**).

Minimum loan rate spread

1.0/1.0 point (graded)

1.0 正确

1.0

Maximum loan rate spread

1.0/1.0 point (graded)

99.0 正确

99.0

检查

Mean loan rate spread

1.0/1.0 point (graded)

2.0 正确

2.0

检查

Median loan rate spread

1.0/1.0 point (graded)

1.0 正确

1.0

检查

Standard deviation of loan rate spread

1.0/1.0 point (graded)

1.7 正确

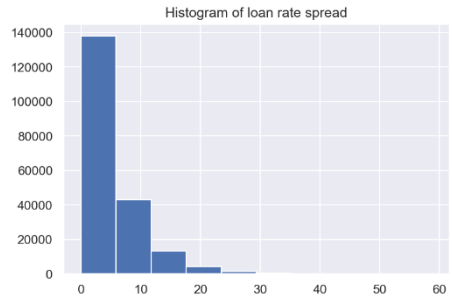
1.7

检查

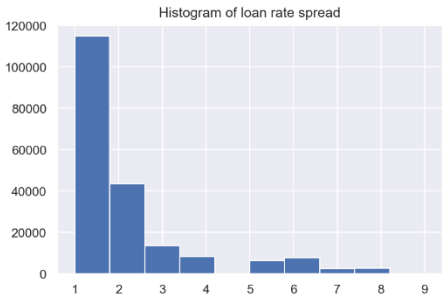
Distribution of loan rate spread

1/1 point (graded)

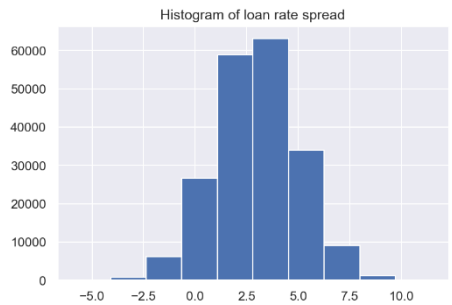
Which of these histograms most closely resembles the distribution of rate_spread?



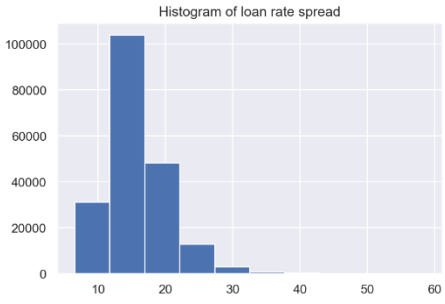
Histogram A



Histogram B



Histogram C



Histogram D

- Histogram A
- Histogram B
- Histogram C
- Histogram D

正确

检查

Loan rate spread across ethnicity and gender

2.0/2.0 points (graded)

Which **two** of the following statements are true?

- [x] Applicants where applicant_ethnicity=3 have a higher rate spread on average than where applicant_ethnicity=1.
- [ ] Applicants where applicant_ethnicity=3 have a lower rate spread on average than where applicant_ethnicity=1.
- [x] Applicants where applicant_sex=1 have a lower rate spread on average than where applicant_sex=2.
- [ ] Applicants where applicant_sex=1 have a higher rate spread on average than where applicant_sex=2.

正确

检查

Applicant income and loan amount

2.0/2.0 points (graded)

For applicants in state 43, which of the following best describes the relationship between applicant income and loan amount?

- (•) A — A higher applicant income is associated with a higher loan amount, on average.
- ( ) B — A higher applicant income is associated with a lower loan amount, on average.
- ( ) C — There is not a strong and obvious correlation between applicant income and loan amount, on average.

正确

检查

Loan rate spreads across counties

2.0/2.0 points (graded)

Limiting just to state 48 and ignoring where county is missing (missing value being -1), which of the following statements is true?

- ( ) A — In state 48, the average rate spread across counties varies substantially, ranging from 0% to 10%.
- (•) B — In state 48, the average rate spread across counties varies substantially, ranging from around 1% to around 7%.

C — Counties within state 48 all have similar levels of loan rates.

正确

检查

Loan types across states

2.0/2.0 points (graded)

Looking just at **states 2 and 3** and just **loan types 1, 2, and 3** which of the following statements are true? Select all that are true.

- [x] A — For loan types 1 , 2 , and 3 , the average rate spread in state 2 is higher than the overall rate among states 2 and 3 .
- [x] B — For loan types 1 , 2 , and 3 , the average rate spread in state 3 is lower than the overall rate among states 2 and 3
- [x] C — For loan types 1 , 2 , and 3 , each average rate spread per loan type in state 2 is higher than for the corresponding loan type in state 3 .
- [ ] D — For loan types 1 , 2 , and 3 , each average rate spread per loan type in state 2 is lower than for the corresponding loan type in state 3 .

正确

检查

Ethics question: 1

2.0/2.0 points (graded)

Data ethics: With great power comes great responsibility. Part of the responsibilities of a data scientist includes thinking about the ethical implications of your work, such as was discussed during the MPP course. It is therefore important to be able to spot when different ethical issues may arise. For the following scenario, familiarize yourself with, then use this Data Science Ethics Checklist, (part of the open source deon command line tool) to find the most relevant ethical concerns.

A bank wants to improve its creditworthiness assessment and decides to hire a team of data scientists to build an algorithm to predict the likelihood that an applicant will default on their loan. The data science team finds that whether or not a loan applicant graduated from a highly selective college is a good predictor of loan default. If this feature is used in the bank's algorithm, which ethical concern is most salient?

- (•) D1: Proxy discrimination
- ( ) C3: Honest representation
- ( ) E4: Unintended use
- ( ) B1: Data security

正确

检查

Ethics question: 2

2.0/2.0 points (graded)

For the following scenario, familiarize yourself with, then use this Data Science Ethics Checklist, (part of the open source deon command line tool) to choose the best answer.

The data science team decides to bring in social media data as they have found that the creditworthiness of an applicant's friends is a good indicator of the applicant's own creditworthiness. Which set of ethical concerns is most relevant here?

- C5: Auditability, E3: Concept drift
- E4: Unintended use, D3: Metric selection
- C2: Dataset bias, D2: Fairness across groups

正确

检查

Ethics question: 3

2.0/2.0 points (graded)

For the following scenario, familiarize yourself with, then use this Data Science Ethics Checklist, (part of the open source deon command line tool) to choose the best answer.

The data science team has addressed the earlier ethical issues in their analysis and has built their algorithm into an application, which takes in an applicant's data and outputs the estimated likelihood of default. The bank rolls out this application across various bank branches. Which set of ethical concerns is most relevant at this phase?

- C1: Missing perspectives, C2: Dataset bias, C3: Honest representation
- E1: Redress, E2: Roll back, E3: Concept drift
- A1: Informed consent, A2: Collection bias, A3: Limit PII exposure