Data Ventures Workshops: Lab 1

Spring 2018

1 Lab 1

This document contains the assignment for lab 1, for those of you that prefer not to work with iPython notebooks.

2 Reading

Look through the lab1preview notebook on github.

3 Download Data

https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv

4 Problem

Now, let's try and use the Pandas' Dataframe structure to analyze a dataset from the Lending Club. The [Lending Club](https://www.lendingclub.com/) is a peer-to-peer lending site where members make loans to each other. For this exercise, we'll be using a public dataset that is generously hosted by Spark. This notebook can be downloaded (with associated data) from its repo.

Any good statistical analysis should start with a fundamental question. Since we are working with a dataset of loans, what interesting questions would we like the answer to? Some options include:

- How likely a loan is to default?
- How is the interest rate of a loan computed from the various loan variables?
- How likely is an individual, with certain attributes, to get his/her loan approved?

While these are all interesting questions, they are fundamentally different and so require different approaches. Today we will explore the second question further and investigate how the interest rate of a loan can be estimated by various different factors.

5 Process

So we want to explore these data, and try to gain some insights into what might be useful in creating a linear regression model. Most importantly, we want to try and separate out "the noise". To do so, we follow the following steps:

- Browse the data
- Data cleanup
- Visual exploration
- Model derivation

These steps are universal to quality Data Science analysis and should be used for nearly every dataset. Today we will trace through the entire data analysis process so that you have a solid backbone on which you can implement the more advanced models you'll soon learn.

6 Exercise 1

Now that we have an idea of the types of data in loansData we can compute some metrics on the data to start discovering interesting trends. For example, you can sort the data and print out the 5 highest interest rates. Using methods from [the Pandas documentation](http://pandas-docs.github.io/pandas-docs-travis/10min.html), compute five interesting trends in the data. This is an exercise in creatively deciding which aspects of the data might provide insights.

7 Exercise 2

Print out one data entry for each of the following errors:

- Outlier values
- Entries with "NA" values
- Incorrect data types

FICO Range is represented as a categorical variable in the data.

We need to change the categorical variable for FICO Range into something numeric so that we can use it in our calculations. As it stands, the values are merely labels, and while they convey meaning to humans, our software can't interpret them as the numbers they really represent.

So as a first step, we convert them from categorical variables to strings. So the abstract entity 735-739 becomes a string "735-739". Then we parse the strings so that a range such as "735-739" gets split into two numbers (735,739). Finally we pick a single number to represent this range. We could choose a midpoint but since the ranges are narrow we can get away with choosing one

of the endpoints as a representative. Here we arbitrarily pick the lower limit and with some imperious hand waving, assert that it is not going to make a major difference to the outcome. In a further flourish of imperiousness we could declare that "the proof is left as an exercise to the reader". But in reality there is really no such formal "proof" other than trying it out in different ways and convincing oneself. If we wanted to be mathematically conservative we could take the midpoint of the range as a representative and this would satisfy most pointy-haired mathematician bosses that "Data Science Dilbert" might encounter.

To summarize - cleaning our data involves:

- removing % signs from rates
- removing the word "months" from loan length.
- managing outliers - remove such rows in this case
- managing NA - remove such rows in this case

Notes: There is one especially high outlier with monthly income > 100K+.This is likely to be a typo and is removed as a data item. There is also one data item with all N/A - this is also removed.

8 Exercise 3

Perform each of the above manipulations on the dataset:

- remove the '%' suffix from each row
- remove the ' months' suffix from each row
- remove the outlier rows
- remove rows with NA

9 Exercise 4

Create a histogram and a box plot for 2 additional factors that you think might contribute to variations in interest rate, and give 2 sentences for each factor explaining the results you find.