Case Study #1

Below is a data set that represents thousands of loans made through the Lending Club platform, which is a platform that allows individuals to lend to other individuals.

We would like you to perform the following using the language of your choice:

- Describe the dataset and any issues with it.
- Generate a minimum of 5 unique visualizations using the data and write a brief description of your observations. Additionally, all attempts should be made to make the visualizations visually appealing
- Create a feature set and create a model which predicts *interest_rate* using at least 2 algorithms. Describe any data cleansing that must be performed and analysis when examining the data.
- Visualize the test results and propose enhancements to the model, what would you do if you had more time. Also describe assumptions you made and your approach.

Dataset

https://www.openintro.org/data/index.php?data=loans_full_schema

Output

An HTML website hosting all visualizations and documenting all visualizations and descriptions. All code hosted on GitHub for viewing. Please provide URL's to both the output and the GitHub repo.

* If you submit a jupyter notebook, also submit the accompanying python file. You may use python(.py), R, and RMD(knit to HTML) files. Other languages are acceptable as well.

Case Study 1 NOTES:

Using R:

Generate visualizations, group data by relevant levels for ggplot

-Time Series Forecast Graph

-Plot interest rates per demographic or geographical labels

Modeling Objectives: Build accurate model, identify most important variables, test model fit

-Create Random Forest Model (Try gradient boosting model too, but observe if data is overfit)

-Create Elastic Net model or lasso regression model

Use R markdown to format results into HTML