9650 Python for data analytics

Spring 2022

Final Project

The management team at Amalgamated Bank Corp (ABC) has asked you to step in as a data analyst. They heard about your rapidly developing Python skills and would like you to help them out.

They are a bank.  Their business consists of issuing loans.  Their problem is that they know very little about their data. They have a data file that has over 300k loans they have issued. The dataset describes information about each loan.  The dataset contains loan amount issued and many other fields that describe the client’s credit factors, such as whether they own their own home or not, their income, how long it has been since they were delinquent on an account, how many years since they received their first credit line, etc.

The problem is that ABC has no idea who their clients are and what their data can tell them about their business. They have asked for your guidance.  They have provided a data file and a data dictionary.  The data dictionary contains a list of fields in the dataset and an explanation for each one.

(To simplify the project, many irrelevant fields have been removed from the dataset.  Those fields still appear in the dictionary, however).

To help our client, here are the steps to take:

1. **Cleanse**

Look for null data in the table, and use the appropriate strategy to handle null data for each column. Explain why this was the strategy that you used.

1. **Outliers**

Search the data for outliers and remove them. Use the appropriate outlier method(s).  Show all work.

1. **Who are our clients?**

The ABC Management team has no idea who their customers are.  They don’t know anything about the type of employment and age, income of their clients.   They  need to know some macro information about their data.  How much (total) do they have out in loans? How much are in default/late? What percentage of the business are default and late?

Along with the information above, provide three other demographics of their customers.

Remember, we are dealing with senior management. The charts must be readable, meaningful and at a summary level

Use at least one group/by or bin in your analysis

1. **Who defaults or has problems repaying loans?**

Create visualizations to show management qualities specific to those who are in default/late.  Show three meaningful visualizations.

Use at least one group/by or bin in your analysis.

1. **Prepare the data for a regression analysis**

The ABC management team has heard that it might be possible to use their data to make predictions.  They don’t know much about data analytics. We are hoping that an example of what is possible will help them understand better.

Prepare the data file to run with a regression analysis.  Use the techniques and methods discussed in class.

1. **Run the regression analysis**

     Show scoring and a confusion matrix.