11/23/2021

GUIDO rossum

Senior data scientist, credit one

Overview:

As per your request, I have collected key take-aways and lessons learned about data analytics.

Investigate factors that indicate whether a customer will default on their loans:

This task entailed answering the following questions:

**1. “How do you ensure that customers can/will pay their loans?**

**We cannot ensure loans will be paid, but we can indicate which factors increase the chances of certain types of customers will pay their loans.**

**2. “What attributes in the data can we deem to be statistically significant to the problem at hand?”**

**The following customer attributes will are statistically significant to the question: Balance limit, Education, Marriage status, age, payment history.**

**3. “What concrete information can we derive from the data we have?”**

**the majority of clients are women. Men default at a higher rate than women. Single and married people default at roughly the same rate. The level of education does not clearly answer the question, however, the majority of clients have at least a college degree. The majority of defaulted accounts occur in the Balance Limit range of 20,000 to 320,000. Young people default more than the elderly. The majority of customers are in the age range of 22-42.**

**4. “What proven methods can we use to uncover more information and why?”**

**The following Python tools and functions were utilized during Exploratory Data Analysis: Pandas to clean and compute data, Numpy to split, segment, and compute data, Matplotlib and Seaborn to visualize data and create reports, and scikit-learn to build classification and prediction models. All of these methods build on each other and allow us to take one feature and see how that relates to other customer data.**

Conclusion:

Credit One’s data is collected in a SQL database. After pull this data from SQL, I cleaned the data, then converted it to a csv. I further cleaned the data in a CSV. I think it is important to note that the conversion to CSV allows more data manipulation than that of a SQL dataframe allows. I learned that cleaning the data at the beginning of the EDA process is essential to achieving the project goal. When dealing with “age”, it is easier to slice the variables into decades rather than each age. Also, visualizing the data is extremely useful in coming to approximate conclusions. I believe moving forward, one has to further develop models in order to effectively evaluate the EDA.

GRANT JOHNSON