Final project

Lending Club is a peer-to-peer lending company that connects borrowers with investors through an online platform. It serves people who need personal loans ranging from $ 1,000 to $ 40,000. Borrowers receive the full amount of the loan granted, minus the initial fee that is paid to the company. Investors buy promissory notes backed by personal loans and pay the Lending Club a service fee. The Lending Club company provides data on all loans granted through its platform during certain periods.

For the purposes of this project, data on loans granted through the Lending Club in the years 2007-2011 were used. Each loan has an indication of whether it has finally been repaid (Fully Paid or Charged off in the loan\_status column). Your task is to build a classification model that, based on this data, will predict with a certain accuracy whether the potential borrower will repay his debt on the loan taken. The data set is accompanied by a file with a description of all variables and a file "FICO Score ranged.pdf" in which the meaning of one of the columns is described in detail.

Below are presented the individual stages of the analysis, the execution of which is necessary to complete the project, and their scoring:

1. Data Processing
2. EDA, i.e. extensive data mining . Describe the conclusions of each graph, support your hypotheses with statistical tests such as t-test or Chi-square. Additionally, answer the following questions:
   1. How does the FICO score relate to the borrower's likelihood of repaying a loan?
   2. How is the credit age related to the probability of default and is this risk independent or related to the FICO score
   3. How is my home mortgage status related to my likelihood of default?
   4. How is annual income related to the probability of default?
   5. How is your employment history related to the probability of default?
   6. How is the size of the requested loan related to the probability of default?
   7. Feature Engineering - create 20 new variables
3. Modeling
4. Perform data clustering (try several methods for this purpose, at least 3) and check if there are any borrower segments, use appropriate methods to determine the optimal number of clusters
5. Train 5 different models, using a different algorithm for each one, and then compare their operation, take the AUROC score as the model quality assessment metric.
6. Check the operation of the previously used methods on compressed data using PCA, compare the results (AUROC score) with the models trained in the previous section.
7. Build the final model, the AUROC score of which will be> = 80%, remember to select important variables, cross-validation and fine-tune the model parameters, and also think about class balancing.