General Data Science: Final project

We will work on a credit risk dataset. Assume that the bank wants to decide whether an individual is **good** or **bad** for a loan/mortgage. The goal is to help the bank figure this out, based on some historical data.

Download the dataset from here.
The goal is to design the best classifier to predict the Class i.e. target variable in the dataset.

I expect the followings to be done:

1. Explore and understand the features:
    a. see how they are related to the target variable.
    b. make a distinction between **categorical** and **numerical** features.
    c. among categorical features see which ones can be **ordinals**
    d. plot a visualization for the numerical and ordinal variables and see how they are related to the target variable. If needed use **PCA** and plot the first few components. (use `sns. PairGrid`, refer to the notebook of last session)
    e. also have a final t-sne visualization only for numerical and ordinal features

2. Create a sklearn **pipeline** to:
    a. Preprocess variables:
        i. handle missing values,
        ii. do scaling if needed,
        iii. one-hot encoding for categoricals,
        iv. change the ordinals to numericals (`OrdinalEncoder `)
        v. and anything else you see necessary
    b. Grid search for classifiers:
        i. **Logistic regression** (with **Lasso** and **Ridge**)
        ii. RandomForests
        iii. SVM

3. Report on the performance of the best classifier over the **test set**:
    a. Plot the decision boundary over the first two **PCs**
    b. Report the scores and **confusion matrix**

Note 1: all classifiers should be **tuned** (perform a grid search over important parameters). At the end create a table (a pandas df) with each row containing each classifier and each set of params and the final classification score (Accuracy, precision, recall and f1).

Note 2: don't forget to have a proper **train test split** and also **cross validation** in your grid search.

Good luck!