# Credit scoring

## Objective

Predict customers, who will default on a loan.

## Dataset

Dataset contains the following tables:

- loans.csv: a sample of one-month loans

### Loans

Every row in the table is one loan, issued for <b>one month</b>.

Column	Data Type	Description
Customer_WID	Integer	Unique customer identifier
DisbursementDate	Date	Date, when loan was disbursed
Age	Integer	Customer age
CustomerType	Factor	Internal customer type classification (1,2)
SOR	Integer	A metric of customer relationship with the bank
MonthsSinceOpen	Integer	Number of months since first account for the customer
MonthsSinceActive	Integer	Number of months since last activity
FinancialMeasure1	Float	Some measure of financial activity of a customer
FinancialMeasure2	Float	Some measure of financial activity of a customer
FinancialMeasure3	Float	Some measure of financial activity of a customer
FinancialMeasure4	Float	Some measure of financial activity of a customer
CRBScore	Integer	Credit rating bureau score
Amount	Float	Loan amount
Default	Boolean	1 - Defaulted on this loan, 0 - paid

### Business variables

For business model evaluation consider using the following information:

<b>Loan application fee (paid only, when disbursed) </b>: 5% of Amount

<b>Annual interest rate</b>: 13%

When loan defaults, bank loses entire loaned amount.

## Guidelines

This project dataset already has a dependent column you need to predict (Default).

In addition to that dataset contains Credit rating bureau score for the customer. This score is your benchmark, that need to improve with additional data provided.

Since you have customer identifier and date, when loan was issued you can try to build additional features based on previous loans.

For credit scoring it is crucially important to start of with model, that is explainable and robust to outliers. Then you might want to investigate more complicated models, and provide a tradeoff between explain ability and predictive power.

At the end you need to provide estimated monthly revenue by using your credit scoring model (refer to business variables above) and what is the cutoff for loan acceptance.

## Project deliverables

<b>Primary project deliverable is a Jupyter notebook</b> with the following sections:

1. <b>Problem statement</b> - in this section you should describe the business problem, that you are trying to solve.

2. <b>Dataset description</b> - describe the dataset you are working with.

3. <b>Data exploration</b> - a free form section, where you are expected to explore the dataset and try to understand field interactions. Try seeing if fields correlate, if there is a date column how some measures change with time and what can you say about it. Try to slice your dependent variable across different independent variable. Treat outliers.

4. <b>Feature engineering</b> - describe and show code that transforms your original dataset into dataset ready for prediction.

5. <b>Feature selection</b> - describe how you approached feature selection: which features you decided to include in your models and why.

6. <b>Modeling</b> - describe different binary classification models you used for this project. Ensure that you are correctly splitting dataset into train and test.

7. <b>Model evaluation</b> - describe which metrics you used for model comparison and why. Compare models fit in step 6 using technical and business measures (potential revenue lift from using one model instead of another, refer to Business variables section).

8. <b>Recommendation</b> - a small section, where you recommend one of the models built in section 6 and 7 for utilization in production. Make sure to describe why using business evaluation terms.

9. <b>Conclusion</b> - final conclusions. Describe your thoughts about the project, is it worth to put resulting model in production, what additional data could benefit your model.

Secondary deliverable is a zip archive with:

1. Project report

2. requirements.txt file with python package requirements necessary for your project.

3. All accompanying code necessary to run project report jupyter notebook.

4. Data files should also be present, <b>but empty</b>, to singify where you expect your data to be.

## Project evaluation

Project is evaluated using the following criteria:

1. Report fullness - report should fully comply with structure described in Project deliverables.

2. Deliverables fullness - your submission should contain all required deliverables.

3. Jupyter notebook report should be runnable without errors, given that python package requirements are satisfied, and data is present in specified locations.

4. Coding standards: source code files/data files/other files in the archive are logically organized, consistently named. Source code uses consistent naming conventions, shows good code reuse and abstractions.

5. Data exploration section should be comprehensive and easy to follow. Should explore majority of features available in the dataset. Charts should be consistently named and follow the same formatting style.

6. Feature engineering, Feature selection sections should be split into subsections according to logical steps in your process. Feature selection should be based on statistical measures.

7. Modeling and Model evaluation - ensure that your train/test split is valid and does not leak future information into your training dataset. Model evaluation criteria is clearly described and used. Estimate and show how you are handling under/overfitting. If your model uses hyperparameters, show how you estimated their values. Very important to add model evaluation in business terms.

8. Recommendation and Conclusions - clear model recommendations in business terms. If given business variables none of the models give revenue lift, clearly describe so in the recommendations, otherwise show potential increase in revenue.