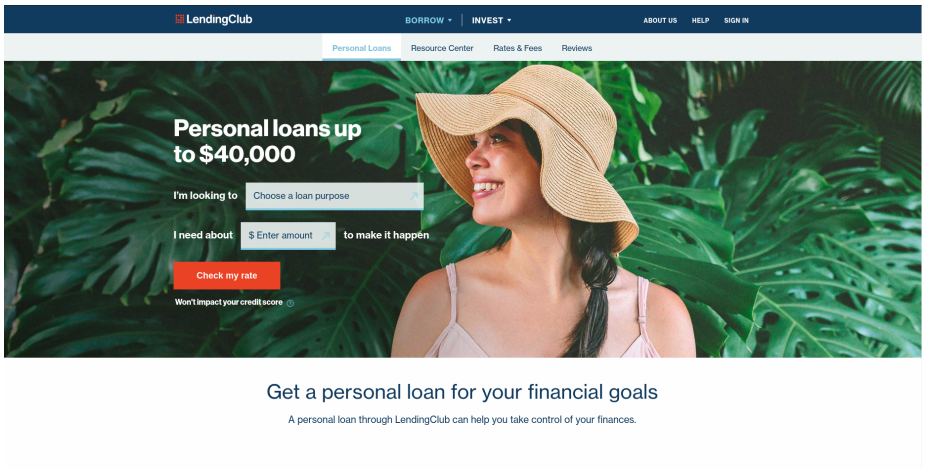

Figure 1: Lending club main

Tabular modeling case

Description

The task is to use a limited version of the openly available Lending Club dataset (provided download link: The dataset) and create a full (but simple) modeling and validation pipeline, using the algorithms and processing steps of your choosing.

The dataset contains information about loan applicants and application-related data. The column to be predicted is called `bad_loan`, and is a binary indicator of whether or not the loan was repaid on time. Focus on the main tasks below, and if time, ambition and your knowledge allows, try attempting one or more of the *additional tasks*.

There are various data types present in the dataset and it is up to you to select variables to use and encoding them properly. Free text fields, for instance, may be omitted altogether or used and analysed with a proper embedding.

As of programming language, Python is preferred, but R is accepted too.

Tasks

1. Load the dataset and describe the dataset briefly with plots and/or summarizing tables of metrics. Give a high level summary of the data with regard to the target variable.
2. Create a preprocessing pipeline and produce a model-ready dataset (based on input requirements of the algorithm you chose).

3. Select variables (based on intuition, as preprocessing or with a importance metrics wrt a modeling technique)
4. Train a model (Suggestion: A Tree-ensemble model such as XGB/LightGBM/Catboost or Neural network)
5. Validate the model performance with regard to predictive performance and generalizability. You're free to choose metrics you find relevant for this particular task.
6. Present code, descriptive analysis and model performance (For instance in a Jupyter notebook).

Additional tasks

1. Include the text columns and see how they affect the model performance. Use a sentence or word encoder. Maybe the text fields need some additional preprocessing.
2. If your algorithm has hyperparameters, reason around good settings or optimize them using suitable method.
3. Produce additional features by some smart transformation of the dataset.
4. Reason around target leaks and whether the feature really was known at the time of the application.
5. Use *Shap* or similar external feature importance package, for feature exploration and selection.

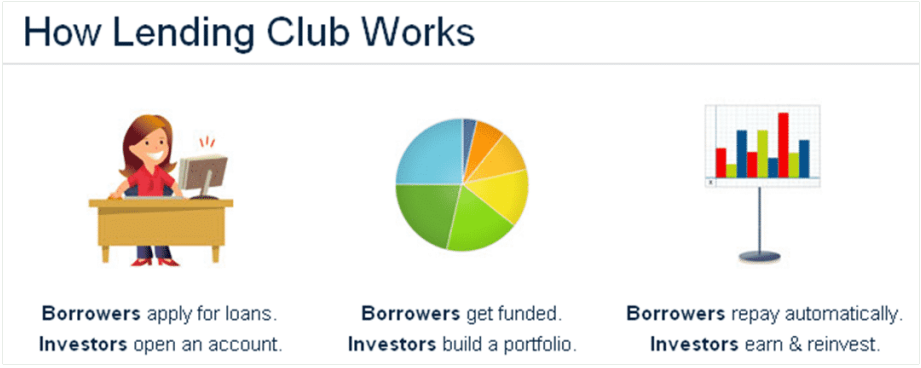

Figure 2: Lending club process