Credit risk assessment using South German credit data

Vahe Charchyan

31.12.2021

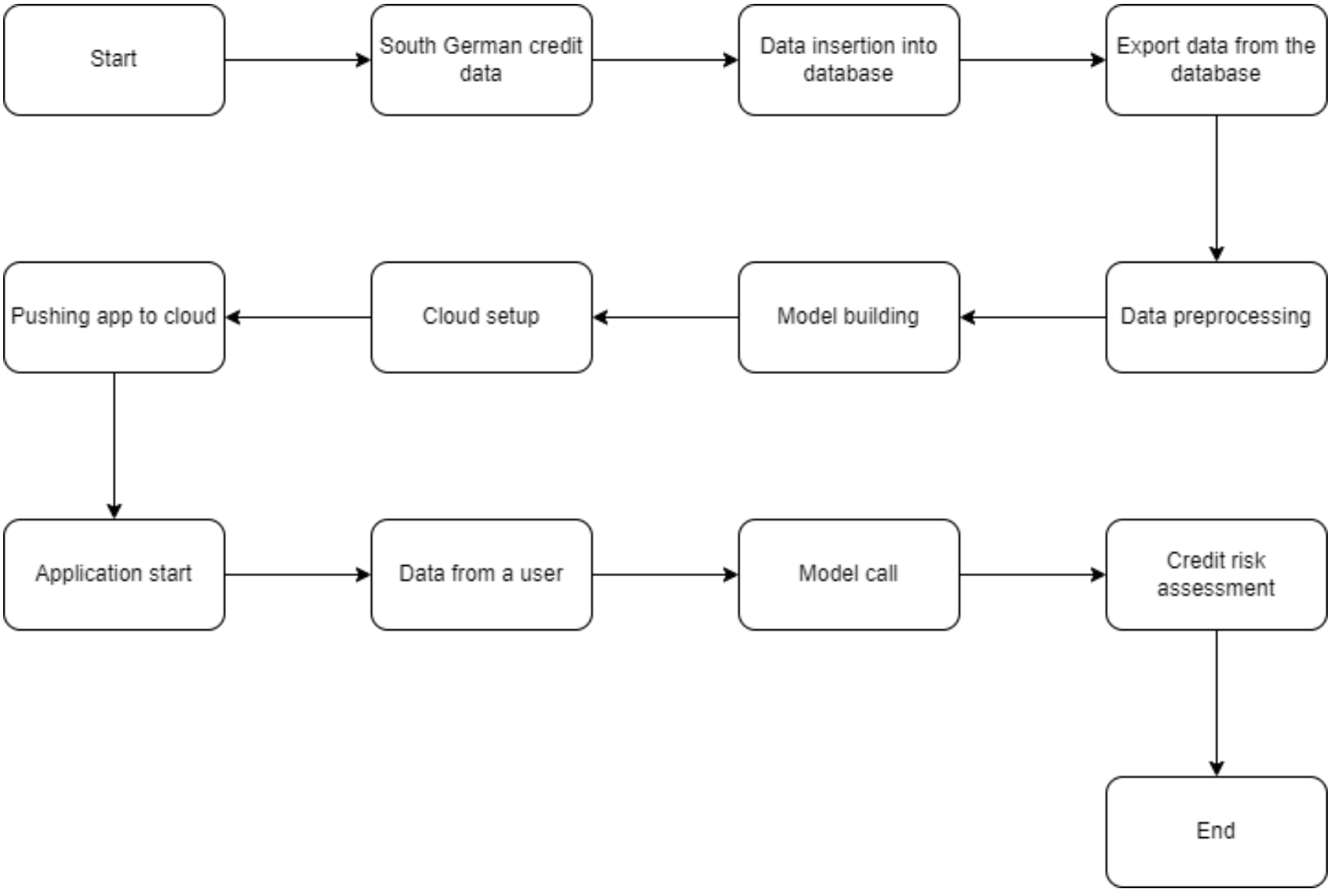
Objective

Developing a credit risk assessment solution to assist credit officers in decision making process. The solution will assess default probability of an applicant and provide possible solution regarding credit issuance.

Benefits

- Detection of defaults
- Credit risk reduction
- Optimal allocation of resources

Architecture



Data insertion into database

- Database creation and connection (if the database exists, open connection to it)
- Import the data

Export data from the database

The data is exported from the database using CQL. Then it is converted to pandas dataframe.

Data preprocessing

- Changing the features' names (actual names are in German)
- Creation of new features (continuous variable discretization, hierarchical clustering based on continuous variables, etc.)
- Transforming continuous variables (applying Box-Cox and MinMax transformation)
- Restructuring the target variable (credit risk) to represent bad credits as 1 and good credits as 0

Model building

Model building includes construction of 3 models (LogisticRegression with onehot encoded categorical variables, LogisticRegression with weight of evidence features, and a simple DecisionTree) and averaging their output to get final result.

Deployment

The model will be deployed to Heroku.

Data from a user

At this stage a user (a credit officer) inserts necessary characteristics of an applicant such as age, credit amount, credit duration, and etc.

Model call

The estimated model will be called to assess credit risk.

Q & A

- What's the source of data ?
  - The data is hosted here.
- What is the type of data ?
  - The data contains both numeric and categorical values.
- What's the complete flow you followed in the project ?
  - Please, look through 4th slide for better understanding.
- How logs are managed ?
  - The logging is applied in model building process almost after each step, but logging results will not be accessible for end users (credit officers). Moreover logging is made accessible via console and a file for developers.

Q & A

- What techniques are used for data pre-processing ?
  - Visualizing relation between target variable and independent variables (features)
  - Box-Cox transformation applied to numeric(continuous) features
  - MinMax scaler applied to numeric(continuous) features
  - Onehot encoding
  - Weight of evidence transformation of categorical features
  - Construction of iteraction features using a simple decision tree
  - Hierarchical clustering

Q & A

- How training was done or what models were used ?
  - We used 3 simple models:
    1. logistic regression where all categorical features are onehot-encoded
    2. logistic regression which takes as inputs only categorical features transformed using the concept of weight of evidence
    3. simple decision tree.
  - After estimating all the models we average their outputs to get final output.
- How prediction was done ?
  - A user provides inputs and clicks on assessment button. The rest happens automatically.
- What are the different stages of deployment ?
  - The deployment was done using Heroku. We pushed code to github and connected Heroku to the github. Then we deployed the project which will update automatically if we update the code on the github.