Improving Decision Making in Credit and Lending

...

Project 2
Rüdiger Hass
Johannes Pastorek
2 July 2020

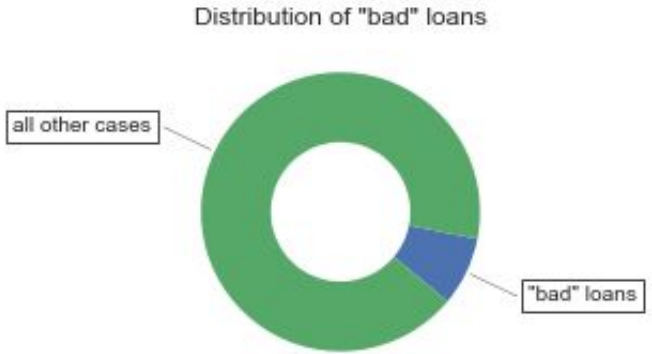
Our Goal

“lower loan risk by identifying patterns from within historical data using machine learning models.”

Historical Data

no. of loans:	365,255
client properties:	123
datapoints:	43,819,365
missing values:	10,605,628
missings in %:	24

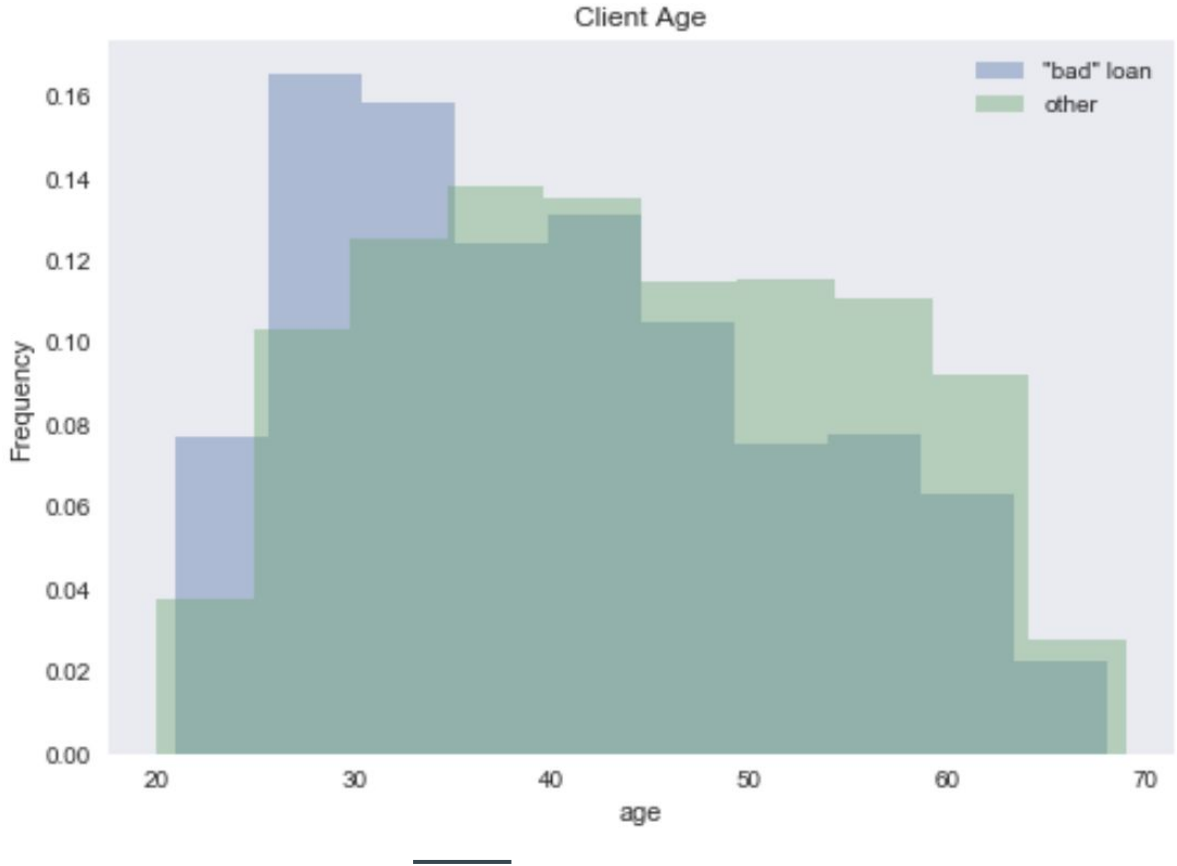
Historical Data



“client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample.”

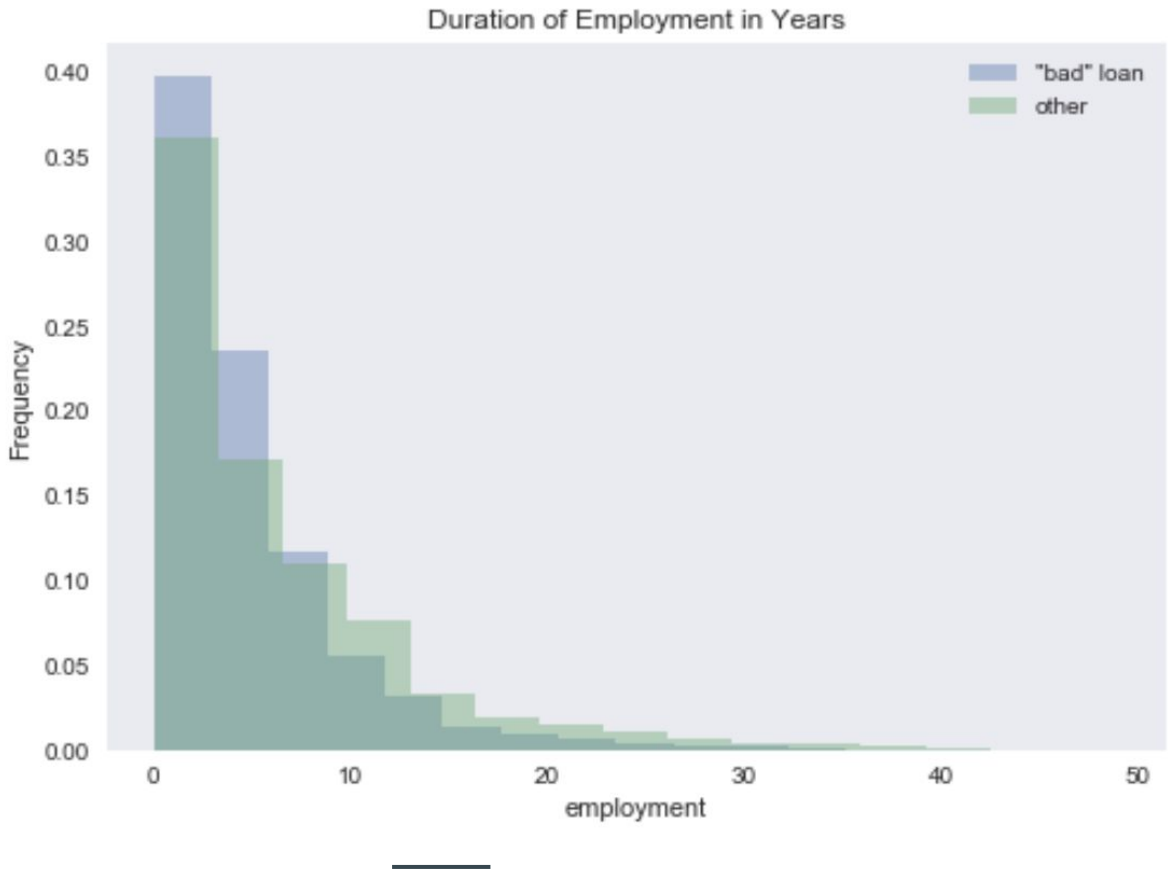
Interesting Features:

- ❑ Age
- ❑ Years employed
- ❑ Gender
- ❑ Education
- ❑ Age of car
- ❑ External Source 1
- ❑ External Source 2
- ❑ External Source 3



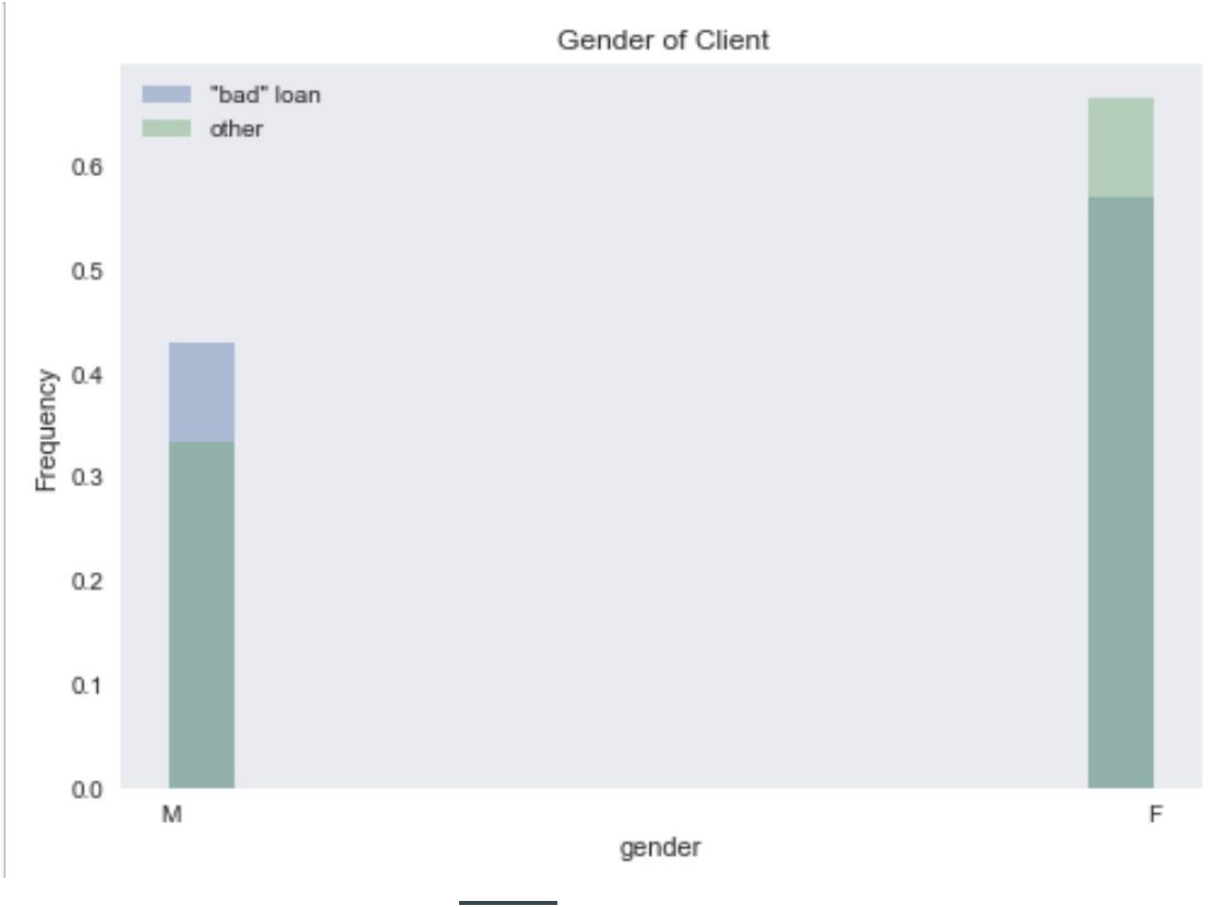
Interesting Features:

- ❑ Age
- ❑ Years employed
- ❑ Gender
- ❑ Education
- ❑ Age of car
- ❑ External Source 1
- ❑ External Source 2
- ❑ External Source 3



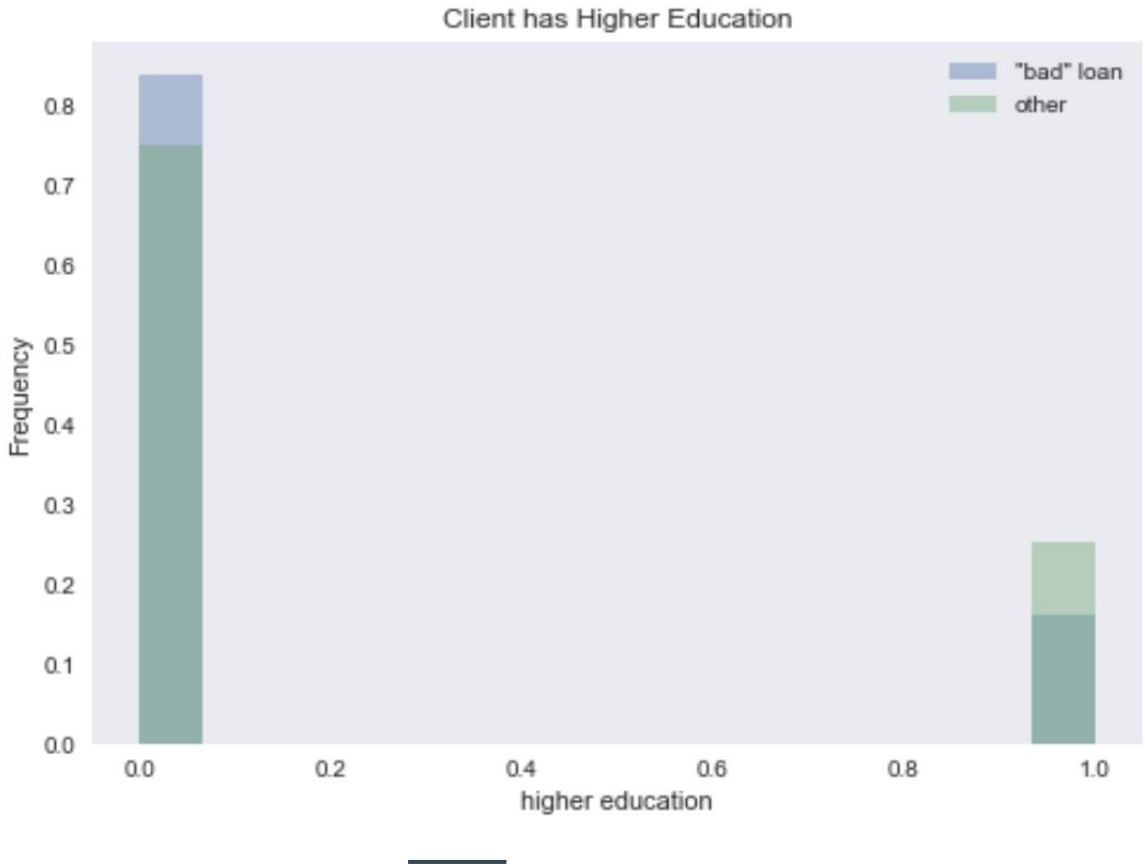
Interesting Features:

- ❑ Age
- ❑ Years employed
- ❑ Gender
- ❑ Education
- ❑ Age of car
- ❑ External Source 1
- ❑ External Source 2
- ❑ External Source 3



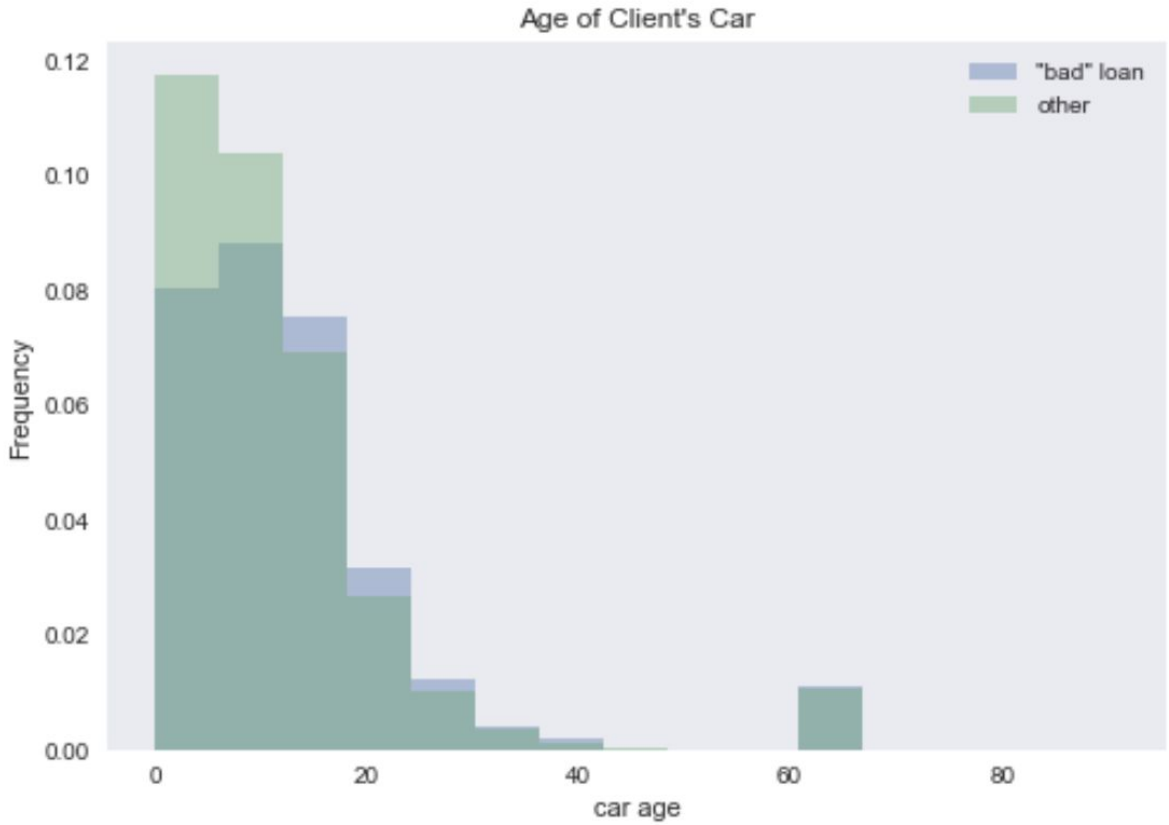
Interesting Features:

- ❑ Age
- ❑ Years employed
- ❑ Gender
- ❑ Education
- ❑ Age of car
- ❑ External Source 1
- ❑ External Source 2
- ❑ External Source 3



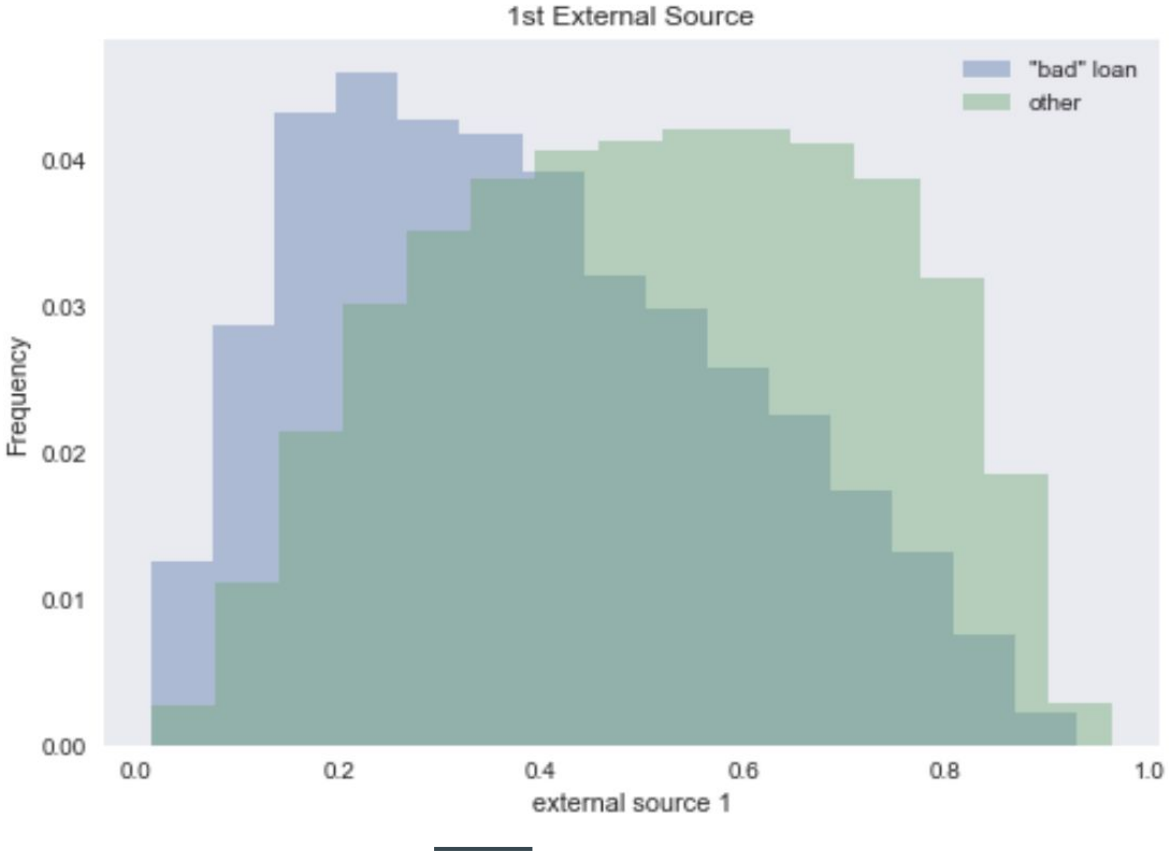
Interesting Features:

- ❑ Age
- ❑ Years employed
- ❑ Gender
- ❑ Education
- ❑ Age of car
- ❑ External Source 1
- ❑ External Source 2
- ❑ External Source 3



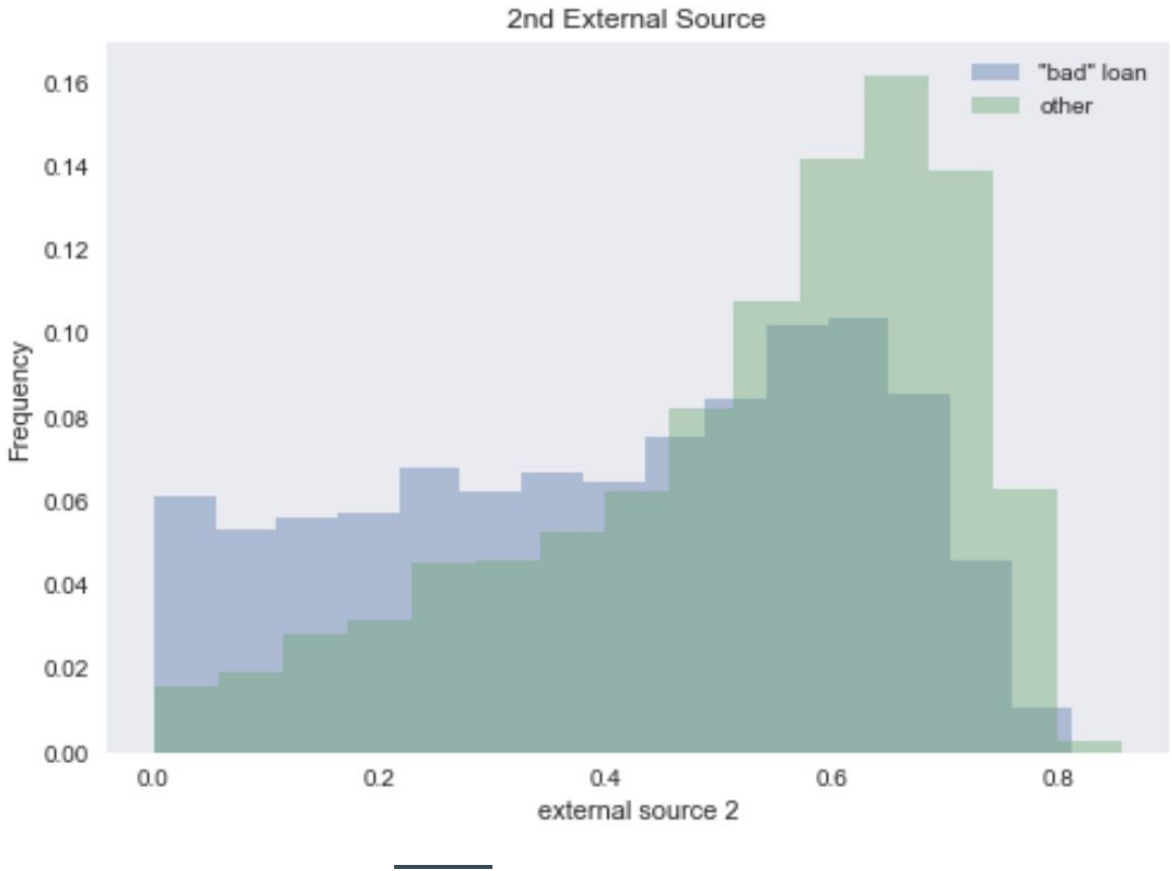
Interesting Features:

- ❑ Age
- ❑ Years employed
- ❑ Gender
- ❑ Education
- ❑ Age of car
- ❑ External Source 1
- ❑ External Source 2
- ❑ External Source 3



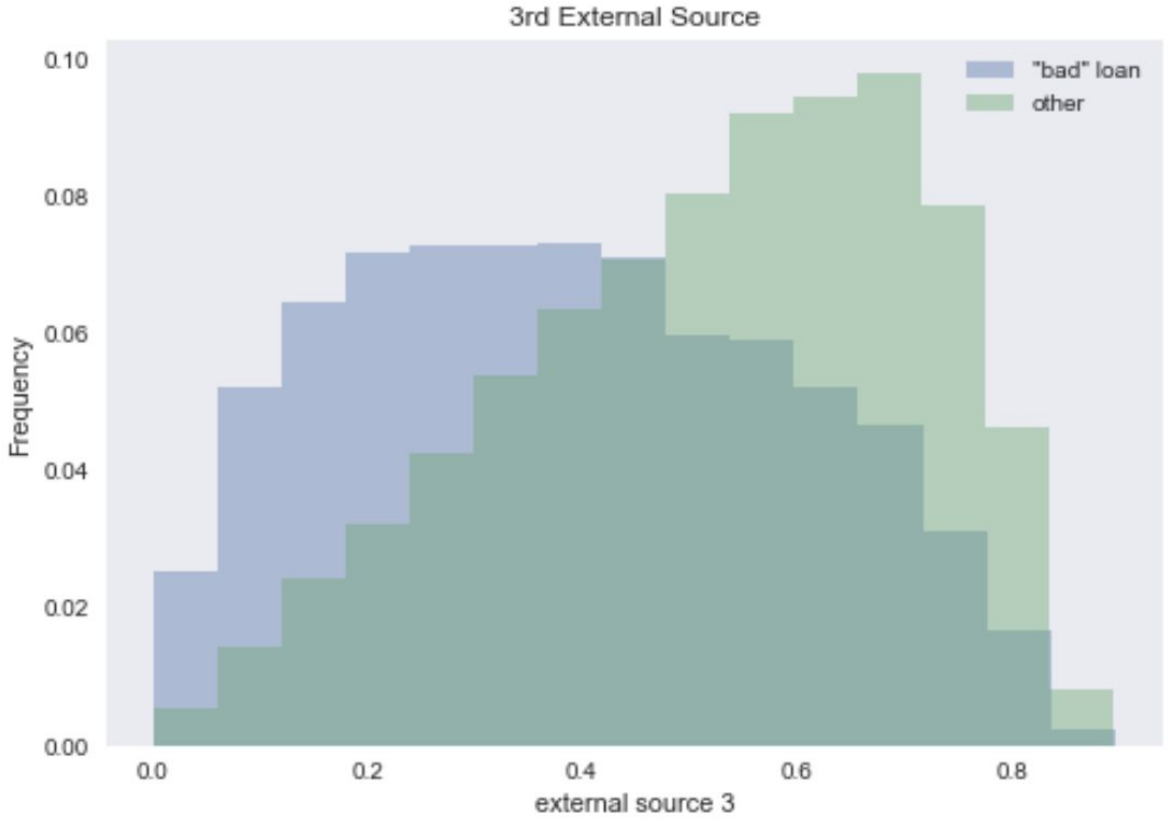
Interesting Features:

- ❑ Age
- ❑ Years employed
- ❑ Gender
- ❑ Education
- ❑ Age of car
- ❑ External Source 1
- ❑ External Source 2
- ❑ External Source 3

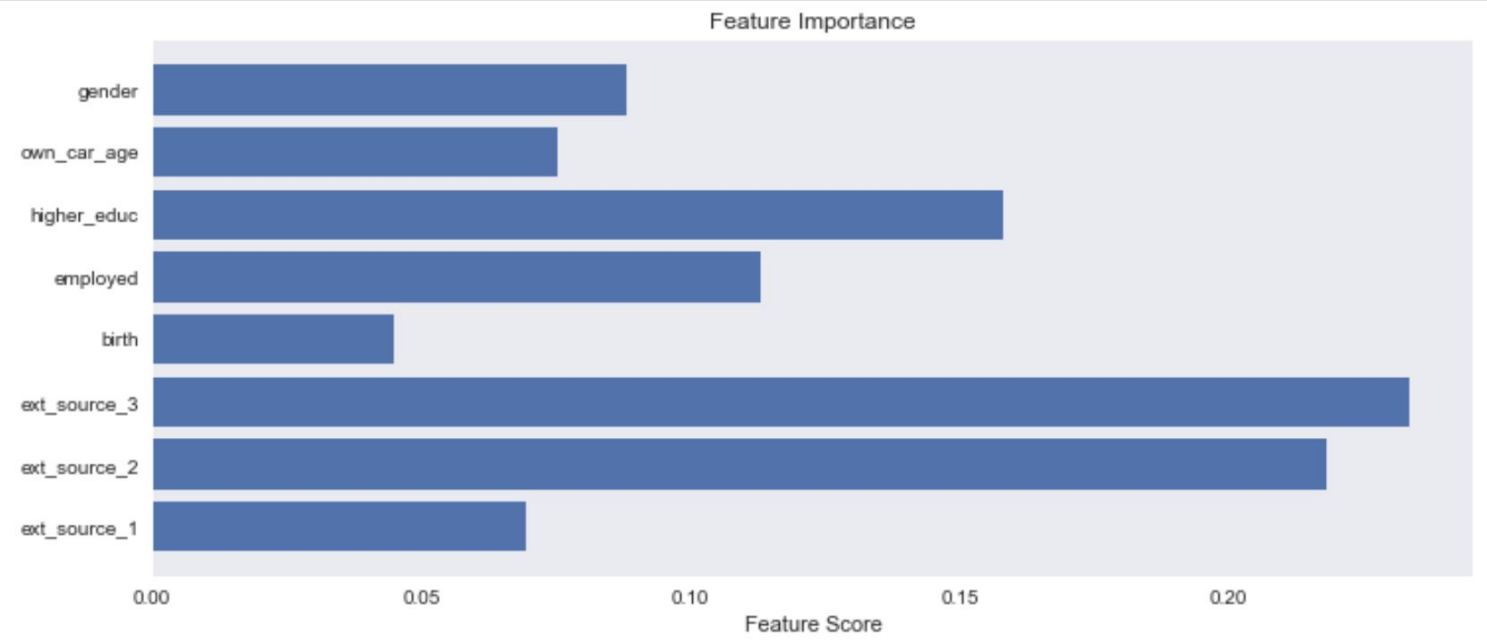


Interesting Features:

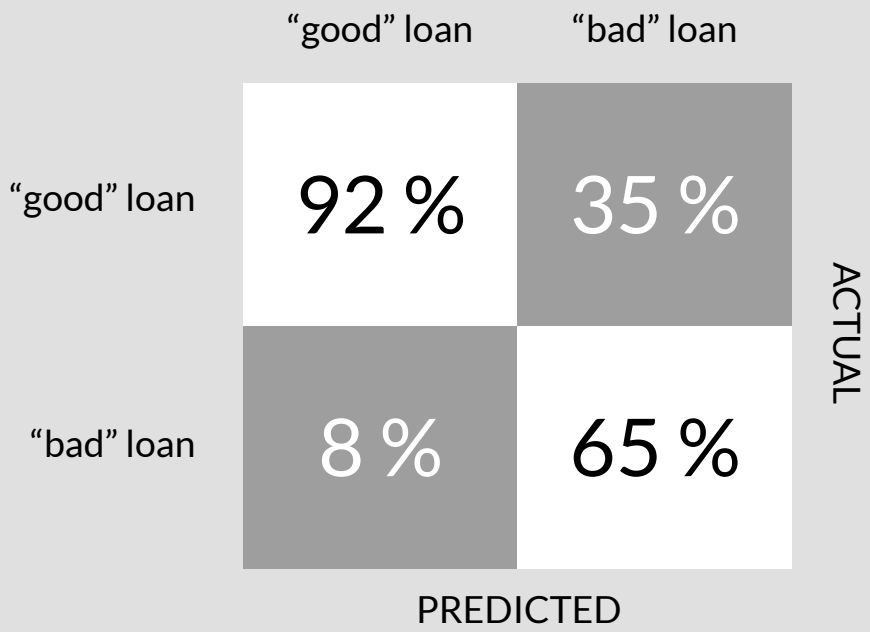
- ❑ Age
- ❑ Years employed
- ❑ Gender
- ❑ Education
- ❑ Age of car
- ❑ External Source 1
- ❑ External Source 2
- ❑ External Source 3



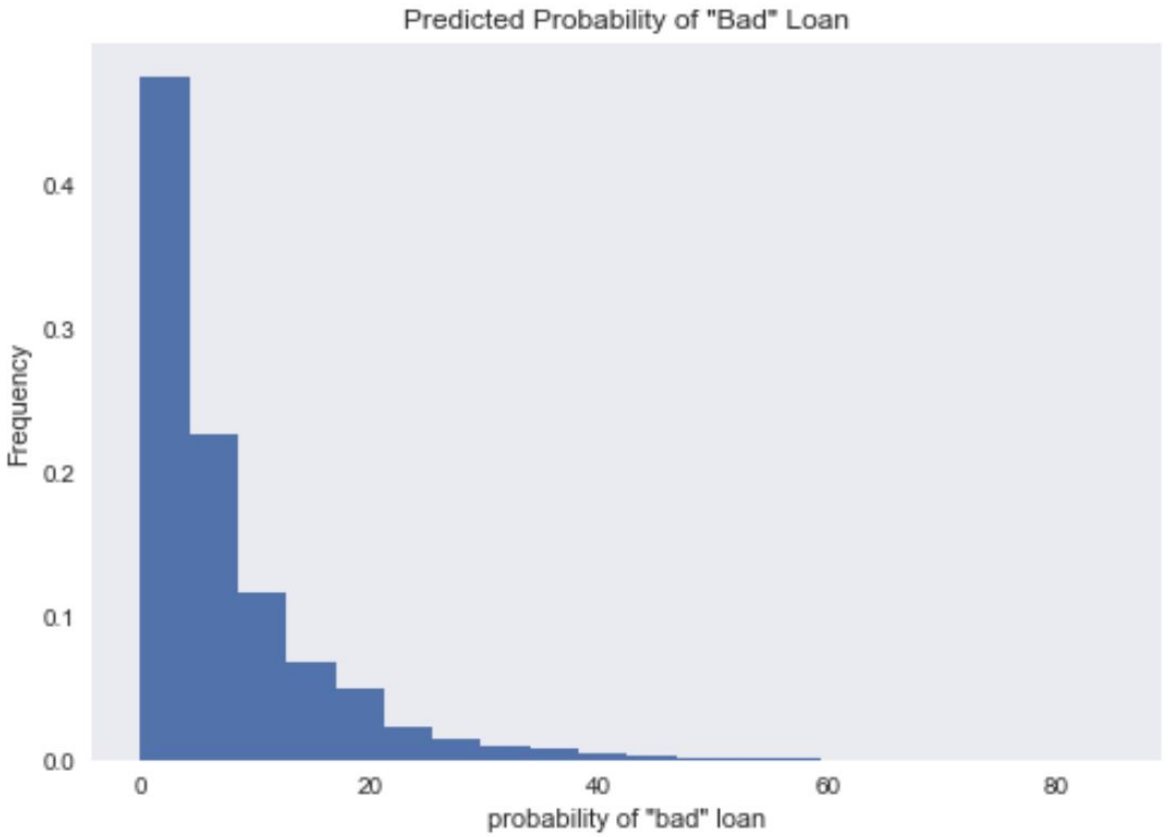
What are the driving factors?



How does the model perform?



Final Prediction

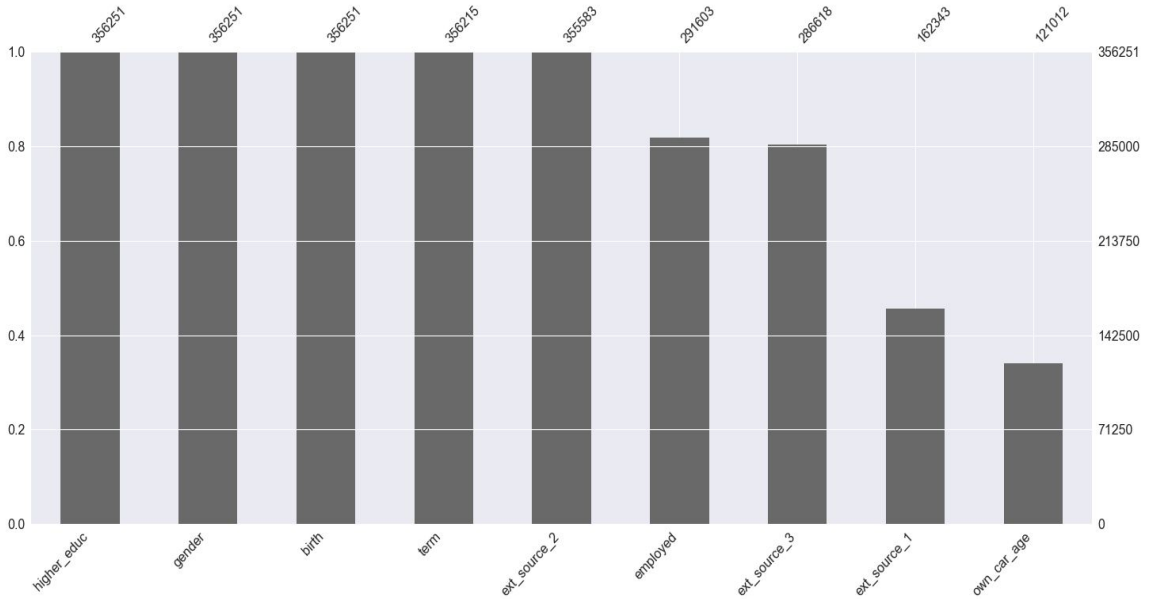


Discussion

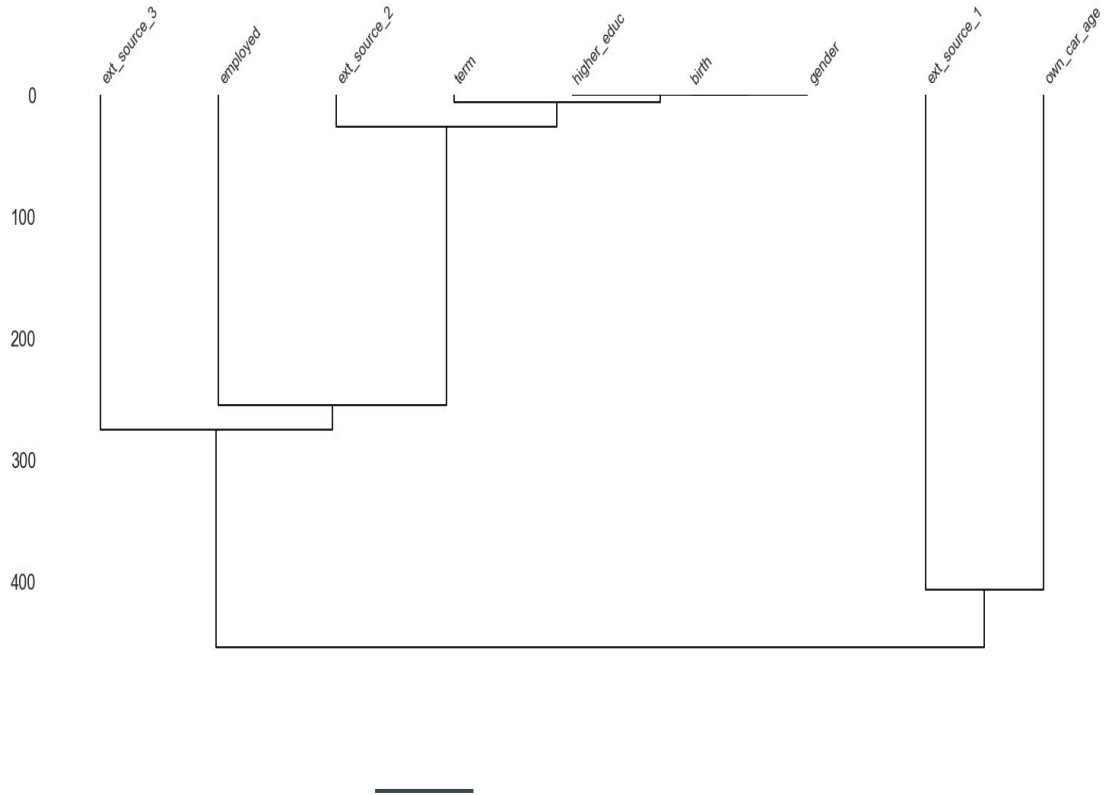
blind spots in data

- Imbalanced dataset
- Missing Values: frequency and distribution
- Traces of multicollinearity
- Little knowledge about features

Missing Values

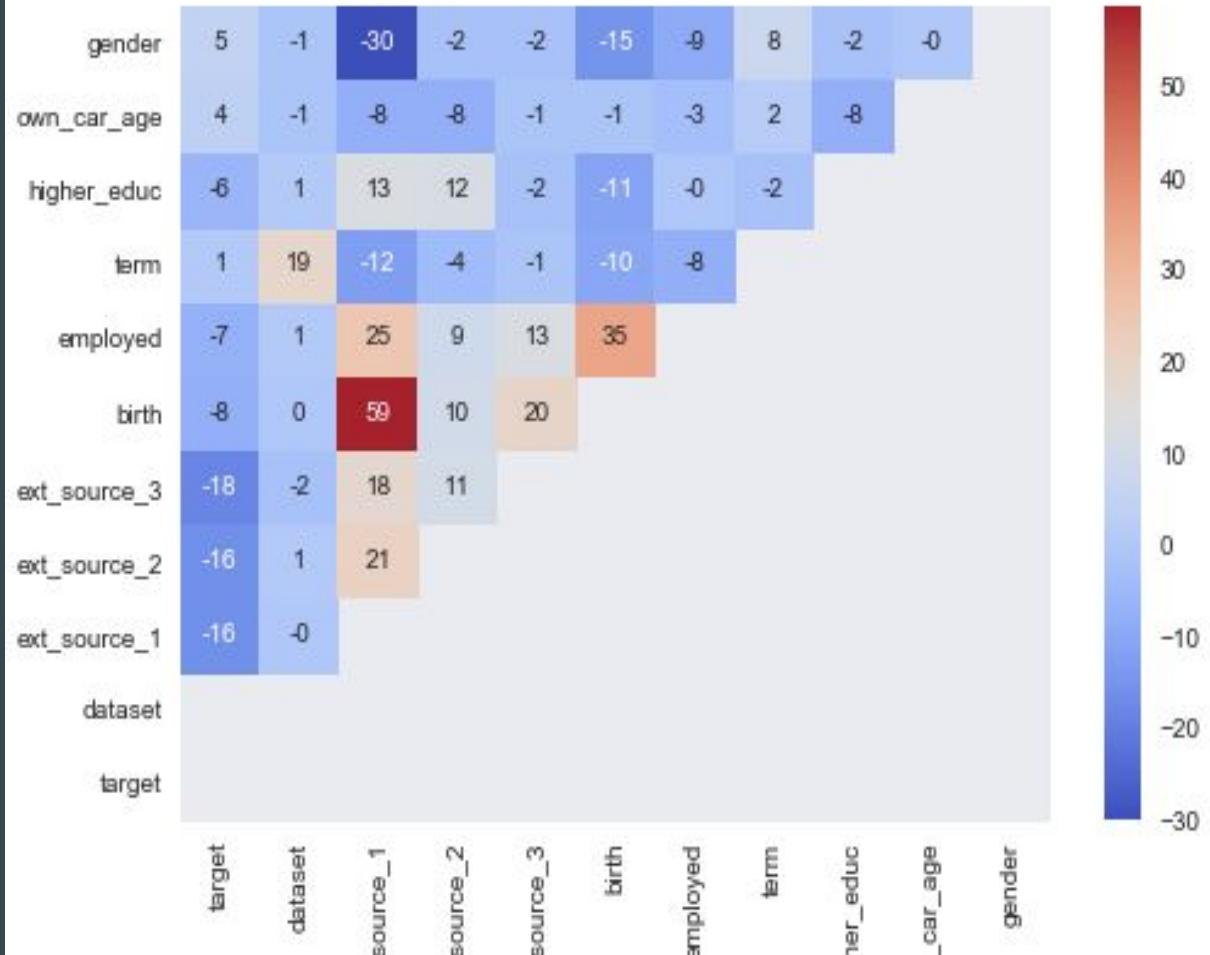


Missing Values



Heat Map

Correlation between variables



Future Work

- ❏ improve missing value handling
- ❏ add more data (external)
- ❏ create more new features
- ❏ improve trade-off scores