DataDriven
PORTFOLIO SOLUTIONS

Peer-To-Peer Lending

GREATYIELDS

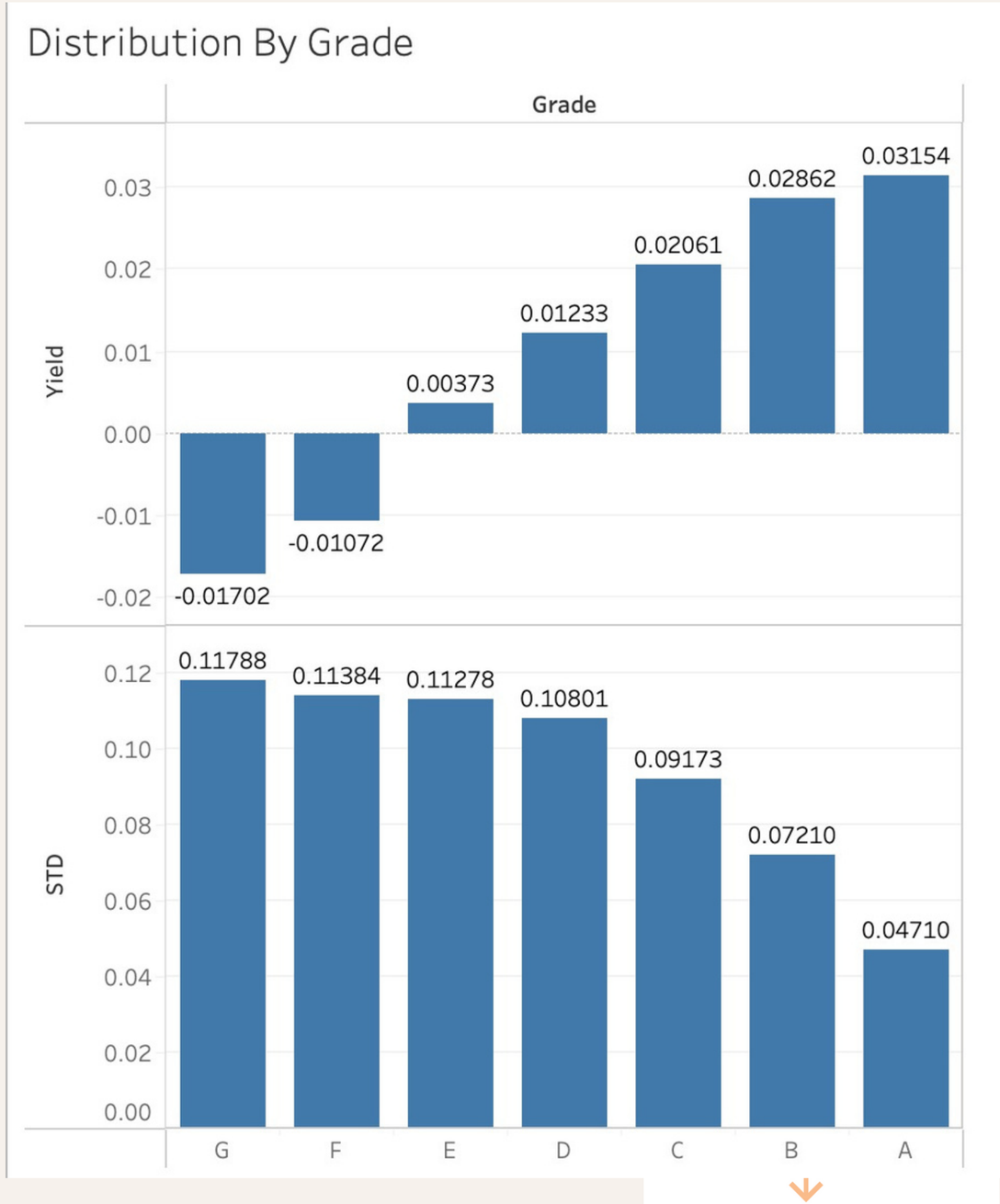
Prepared By: Shelly Levy, Tom Saacks and Or Liberman

date

1 May 2023

Expected Realized Returns and Distribution by Loan Grade

- High-grade loans - lower returns
- High-grade loans - higher STD
- The loans pose significant risk.



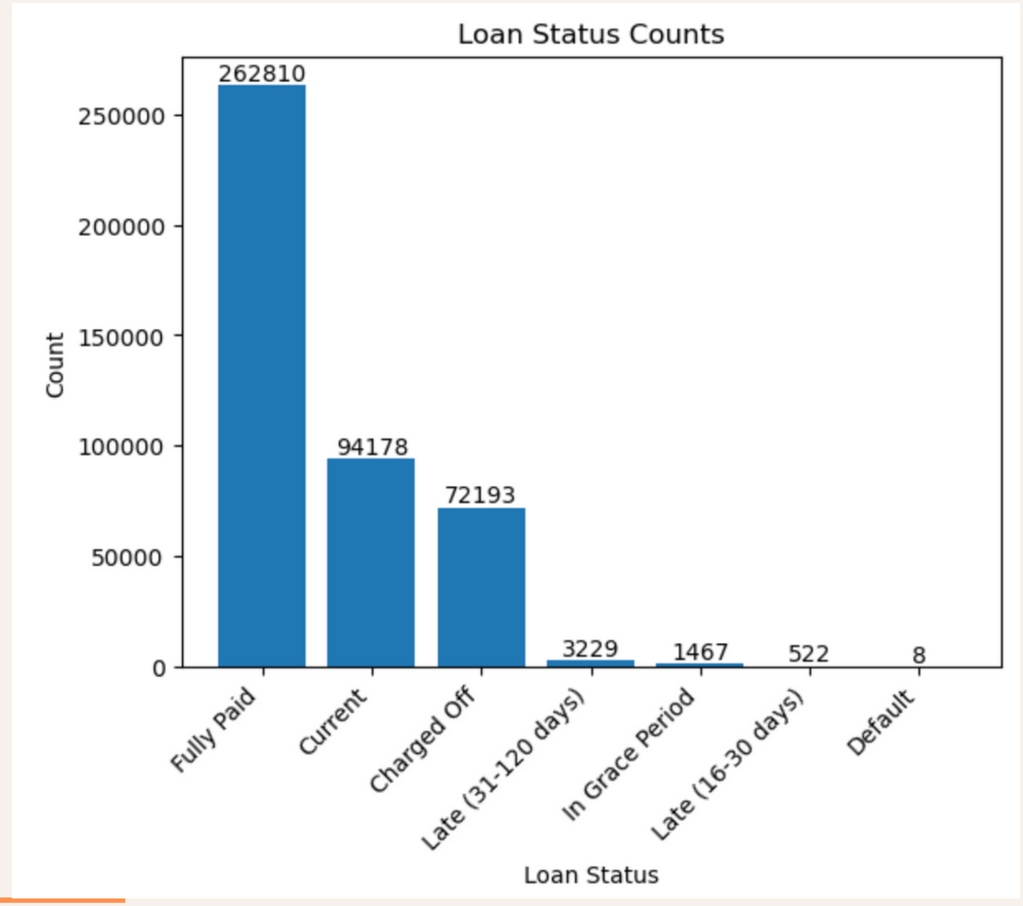
Data Consistency and Validation

- 2016 loans, 8 files, 2018-2019 snapshots by quarter.
  - Consistent loan identification over time.
  - Feature consistency checked except 2018.
  - Leak test for data consistency.
- Data type validation for accurate analysis.

DataDriven
PORTFOLIO SOLUTIONS

Selecting Relevant Loans

- The raw data contains 434407 instances.
- Removed incomplete loan statuses.
- Removing infrequent joint applications.



DataDriven
PORTFOLIO SOLUTIONS

Selecting Relevant Loans

- The raw data contains 434407 instances.
- Removed incomplete loan statuses.
- Removing infrequent joint applications.

Loan Status by Application Type

loan_status application_type	Charged Off	Fully Paid	Total
Individual	98.05%	98.23%	98.19%
Joint App	1.95%	1.77%	1.81%

- 328,956 instances remaining after 19% removal.

DataDriven
PORTFOLIO SOLUTIONS

Target Variables

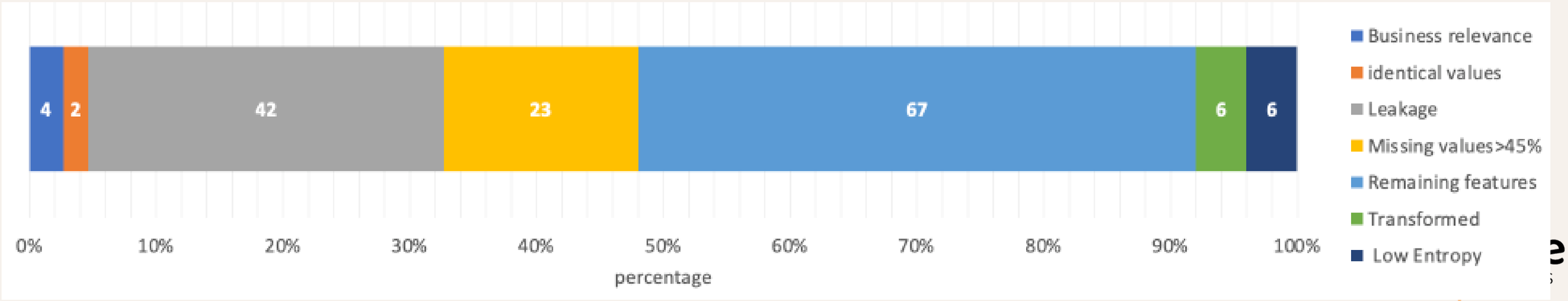
Calculation

- Continuous Variable: Yield
- Binary Variable: Above 2%
- Exploration of Both Variables
- Decision on target variable in 3-4 weeks

DataDriven
PORTFOLIO SOLUTIONS

Feature selection

- The raw data contains the 150 features.
  - Removed leakage features (42)
  - Removed features with >45% NA's (23)
  - Removed business-irrelevant features (4)
  - Evaluated distribution of feature values (12).
  - Remaining features: 66.
- Feature selection ongoing, 2-3 weeks.



Feature selection

Evaluated distribution of feature values

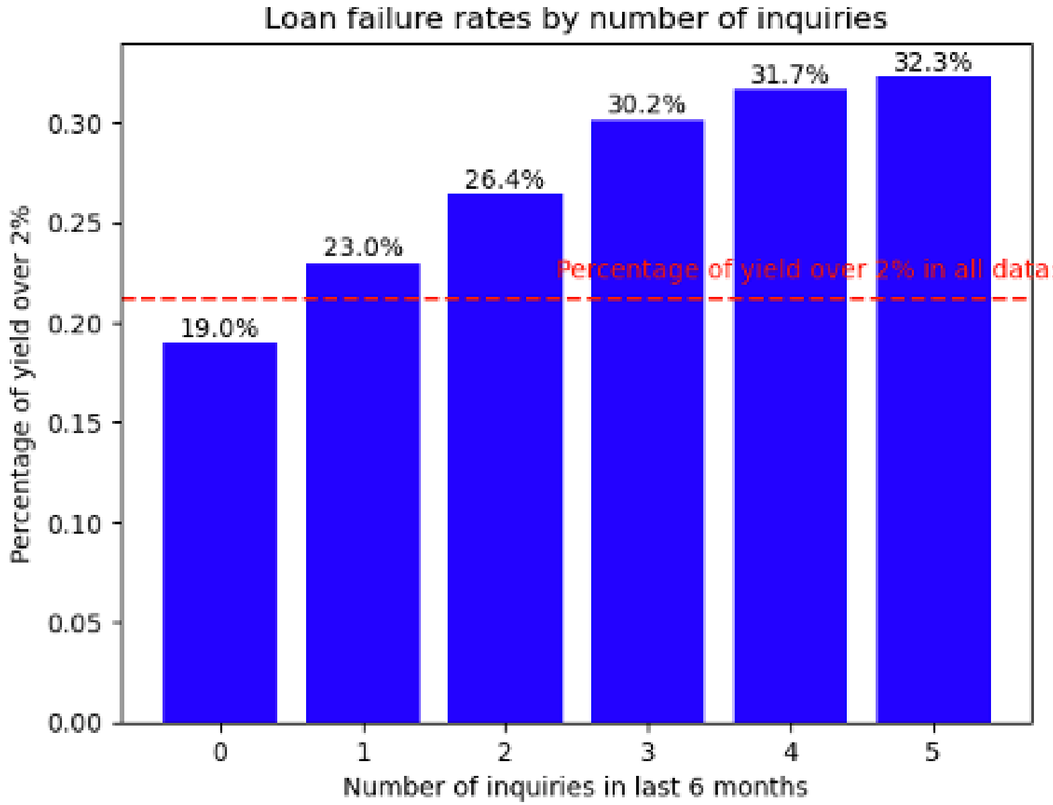
- Removed features with identical values (2).
- Analyzed feature entropy and low frequency values.
- Considered relevance to the problem.
    - Removed/transformed based on feature relevance
- Ongoing evaluation, 2-3 weeks until completion.

DataDriven
PORTFOLIO SOLUTIONS

Feature selection

Evaluated distribution of feature values

- inq_last_6mths -Credit inquiries in 6 months



Number of inquiries	Percentage of data
0.0	60.5%
1.0	26.6%
2.0	8.7%
3.0	3.0%
4.0	0.9%
5.0	0.3%

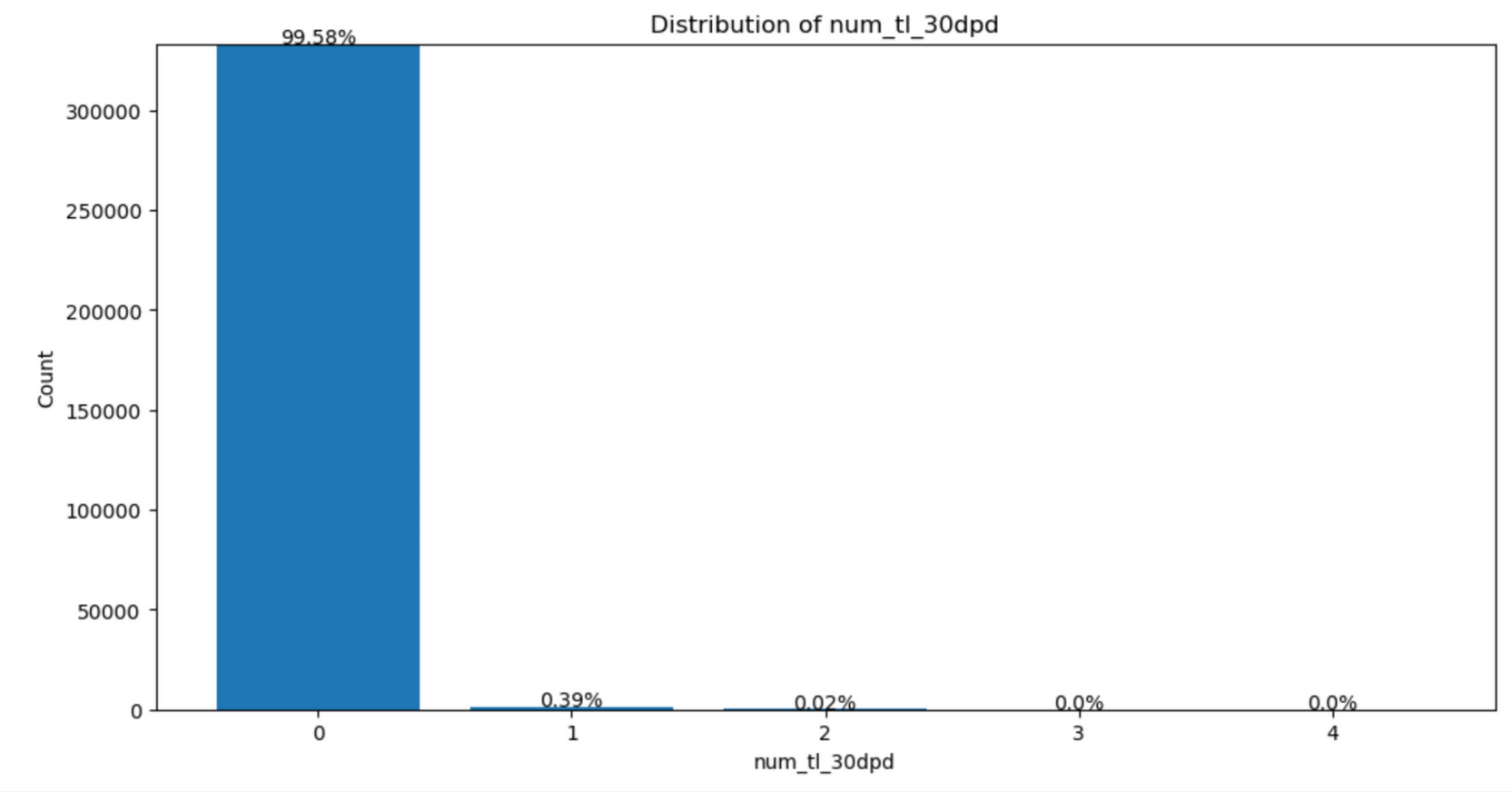
- Feature Transformed

DataDriven
PORTFOLIO SOLUTIONS

Feature selection

Evaluated distribution of feature values

- num_tl_30dpd-Borrower delinquency risk indicator.

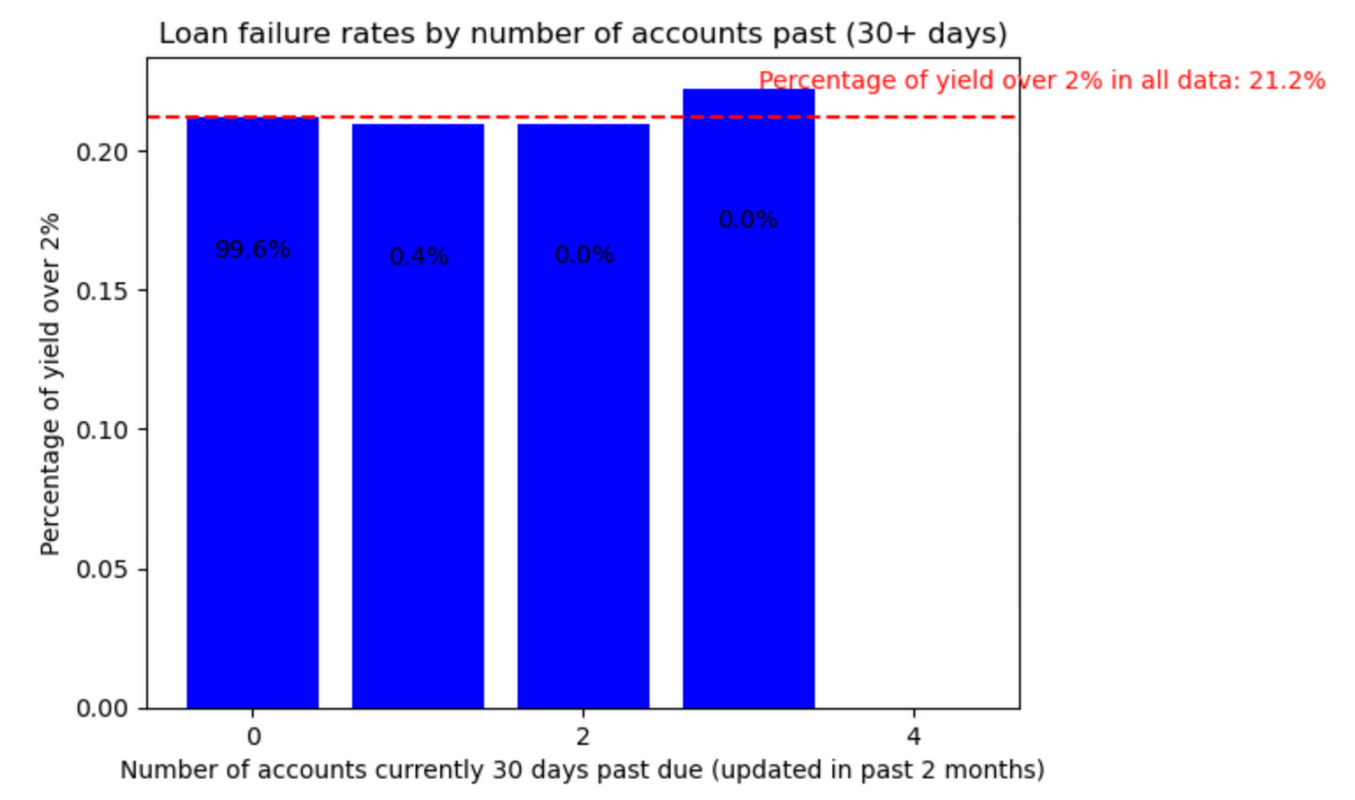


DataDriven
PORTFOLIO SOLUTIONS

Feature selection

Evaluated distribution of feature values

- num_tl_30dpd-Borrower delinquency risk indicator.

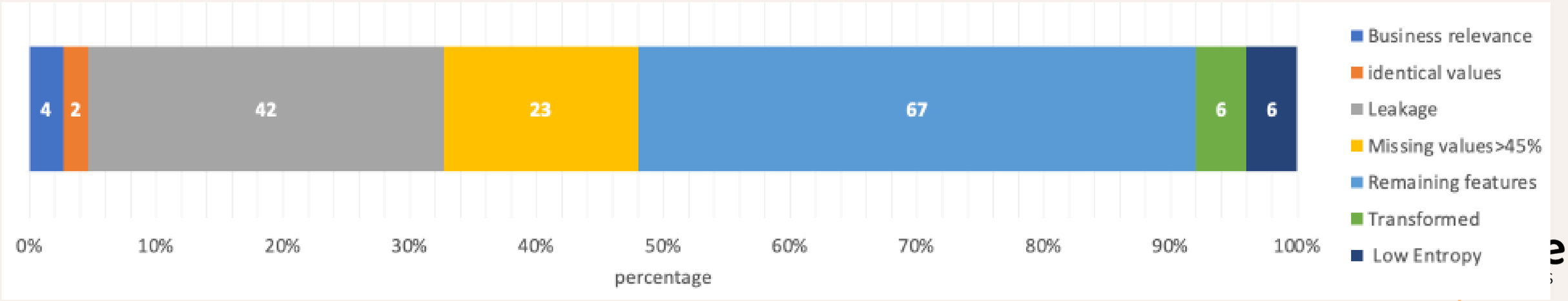


- Feature removed

DataDriven
PORTFOLIO SOLUTIONS

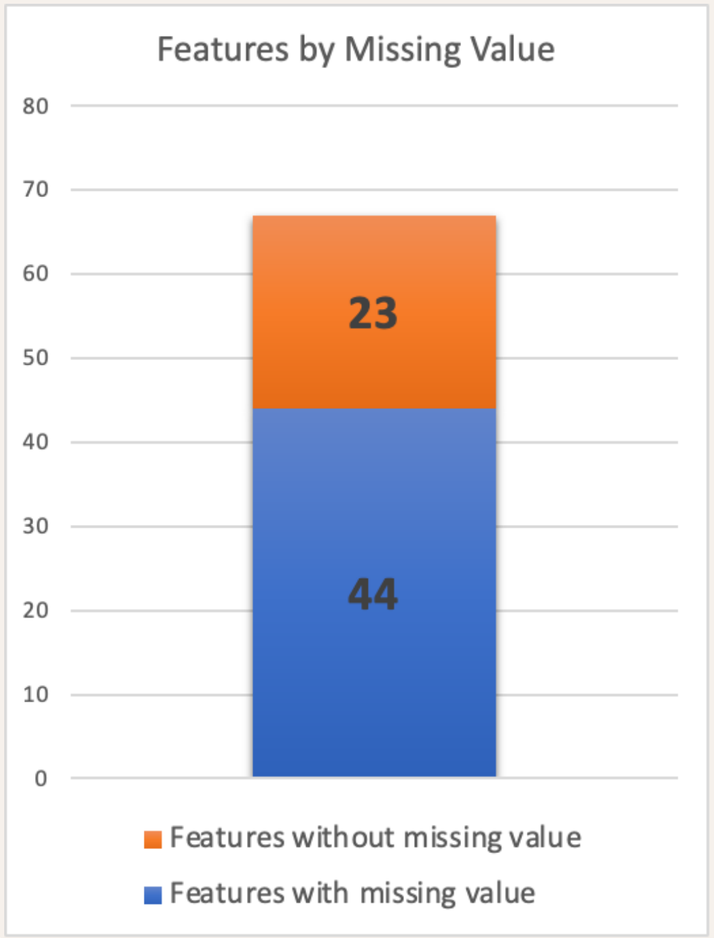
Feature selection

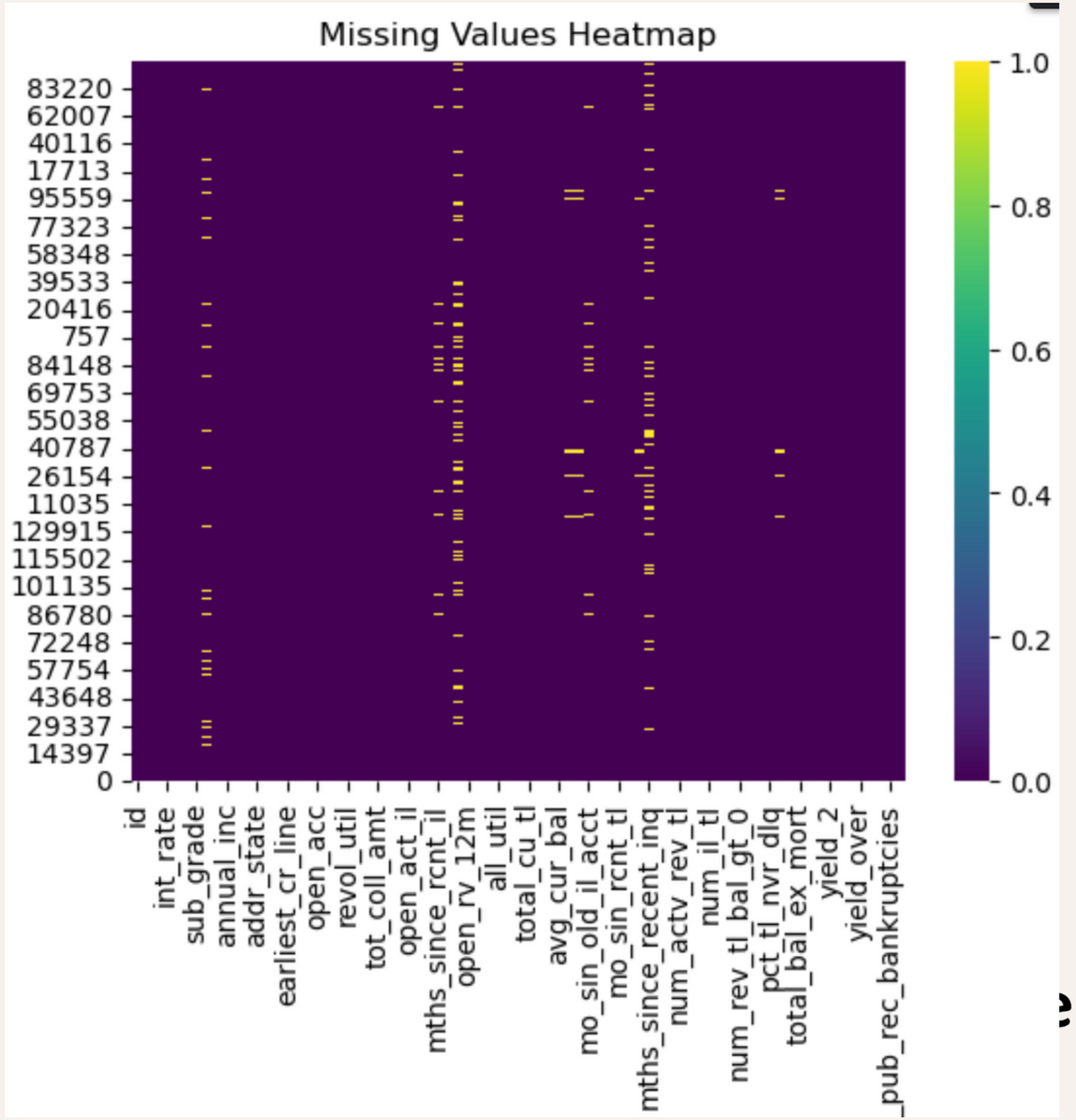
- The raw data contains the 150 features.
  - Removed leakage features (42)
  - Removed features with >45% NA's (23)
  - Removed business-irrelevant features (4)
  - Evaluated distribution of feature values (12).
  - Remaining features: 66.
- Feature selection ongoing, 2-3 weeks.



Missing Data Handling

- Low missing values (<13%) in features.





Missing Data Handling

- Removed instances that <0.06% missing values
- Completion process ongoing for 2 weeks
- Various completion methods were used.

DataDriven
PORTFOLIO SOLUTIONS

Missing Data Handling

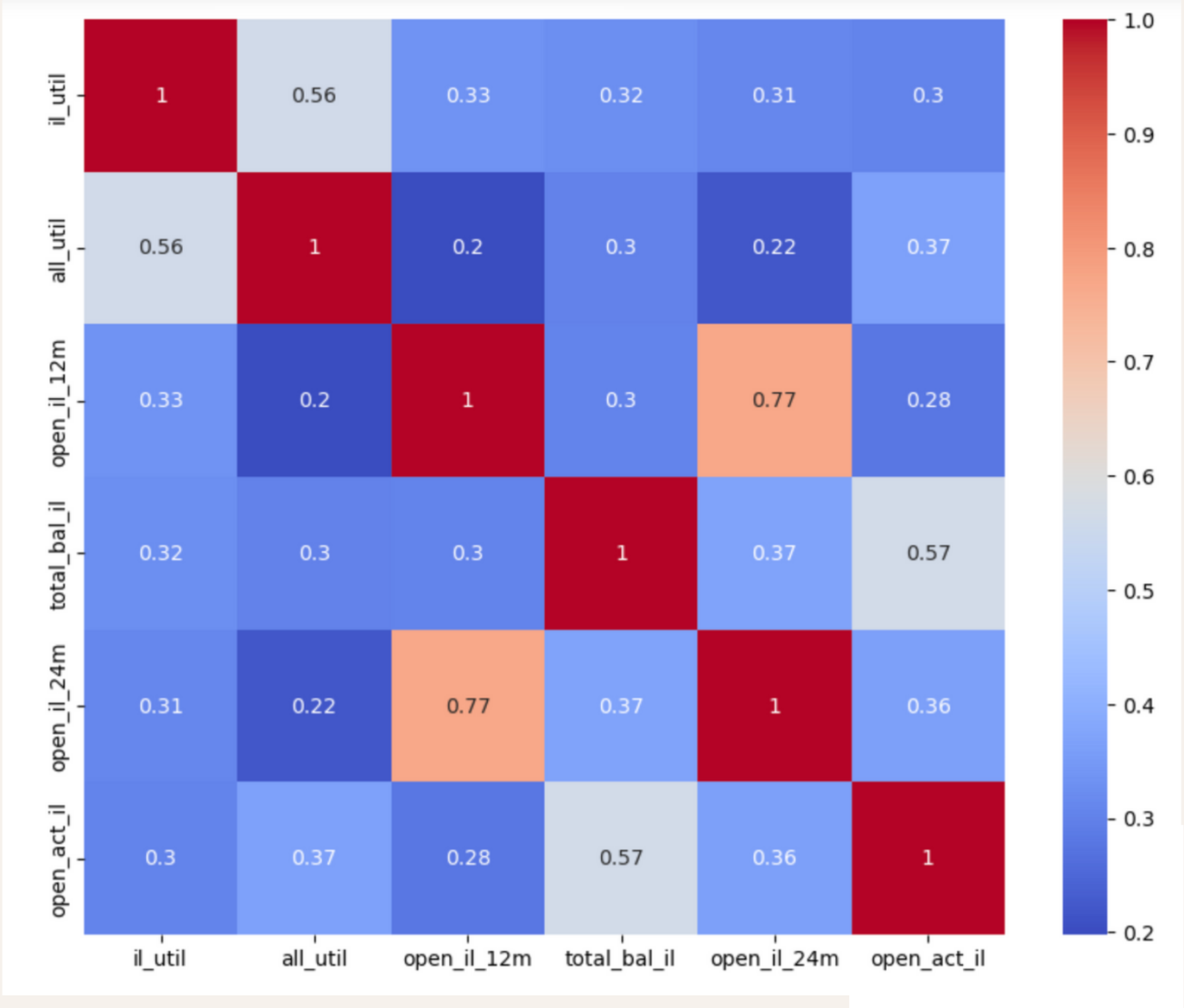
- emp_length- Employment length in years
- 6.51% missing values
- Implement grouping.

```
Home ownership: RENT, Annual income range: (200000.0, 205000.0], Mode emp_length: 10.0
Home ownership: ANY, Annual income range: (45000.0, 50000.0], Mode emp_length: 0.0
Home ownership: ANY, Annual income range: (25000.0, 30000.0], Mode emp_length: 5.0
Home ownership: RENT, Annual income range: (255000.0, 260000.0], Mode emp_length: 4.0
```

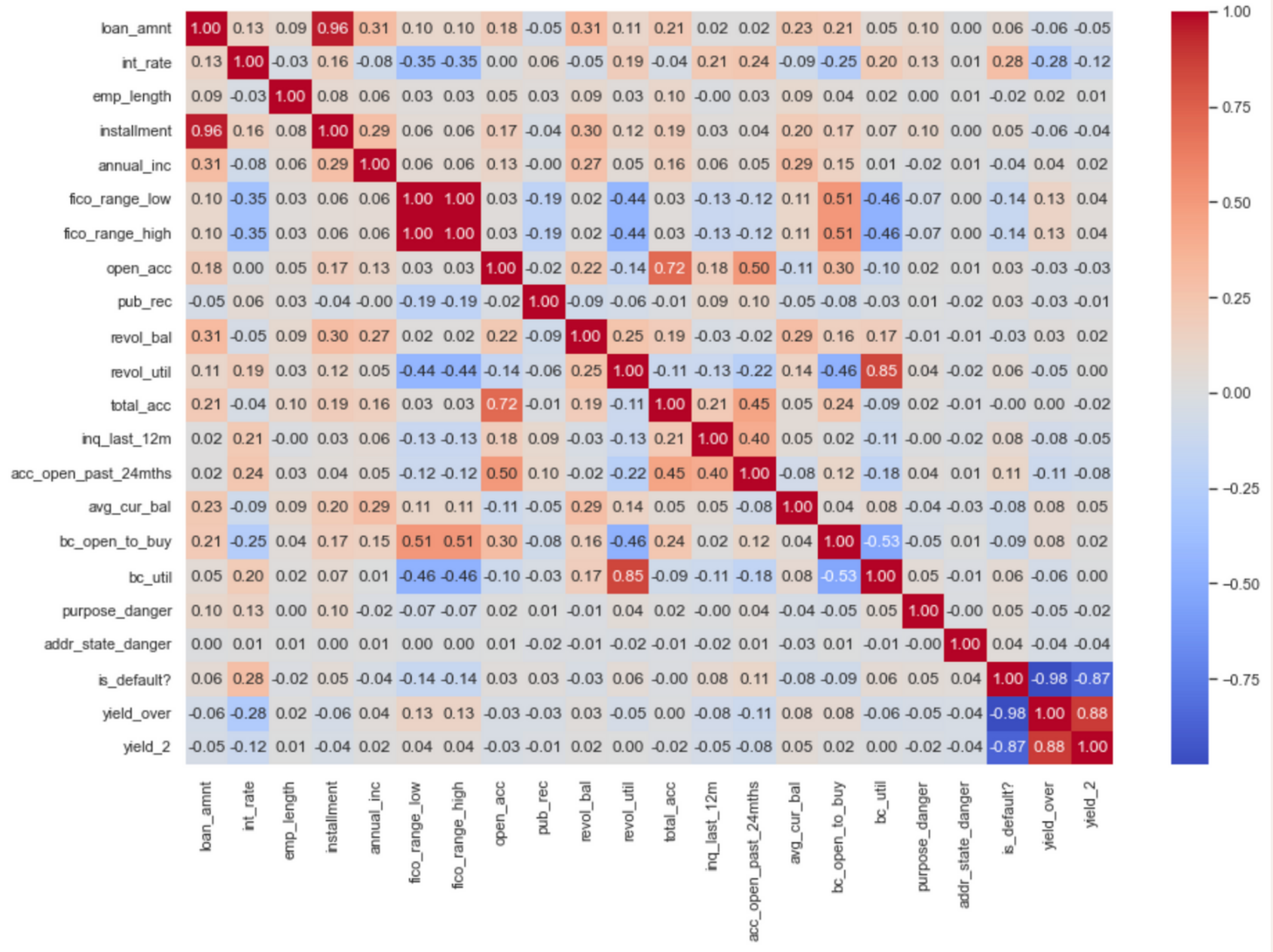
DataDriven
PORTFOLIO SOLUTIONS

Missing Data Handling

- Il_util- the credit utilization on all installment accounts.
- 13.19% missing values
- Implement Linear regression and rounded results.



Correlation HeatMap



DataDriven
PORTFOLIO SOLUTIONS

Potential Pitfalls:

- Biased data from borrower.
- Bias from missing data handling.
- Incorrect transformation methods.
- Overfitting from highly correlated features.
- Two-week estimate for outlier handling.

Next Steps:

- Continuing data preparation.
- Commencing model development.

DataDriven
PORTFOLIO SOLUTIONS

Concluding Remarks and Q&A

THANK **YOU**

DataDriven
PORTFOLIO SOLUTIONS