Predicting the probability of loan default

Nancy Truong, Nicolai Dimkovski Gottschalk

Github repository for the project:
myngoc-trg/BERN02_Project_Credit_Risk: Evaluate predictive models for credit risk classification, identifying whether a borrower is likely to default or not. Techniques: PCA, t-SNE, SMOTE, Random Forest, Neural Network

Introduction: Project and Data Overview

- Loan Default: When a borrower fails to fulfil loan obligations.
- Why important? Impacts lending decisions, profitability.
- Methods: PCA, tSNE, Random Forest Classification, Neural Network
- Credit Risk Dataset: https://www.kaggle.com/datasets/laotse/credit-risk-dataset
  - 11 features: numerical + categorical
    → **17** after one-hot encoding + drop the most dominant category
  - 1 target (**loan_status**): 1 = default, 0 = non-default.
  - 32 581 observations.
    → After excluding duplicates and outliers: 32,409 observations

Data Cleaning

Missing data(NA): **person_emp_length** and **loan_int_rate**

- **Person_emp_length**: Filled with median
  + Checked: **person_age - Person_emp_length** > 15 → none.
- **Loan_int_rate**: Higher grades (A-G) have higher interest rate.
  Fill with calculated mean interest rate per grade.

Imbalance data: more default

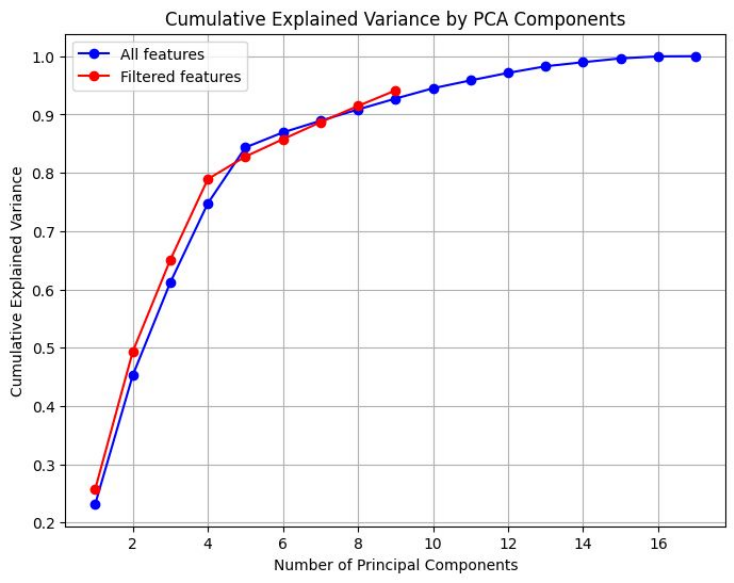
- Use **SMOTE** to create synthetic samples for minority class.

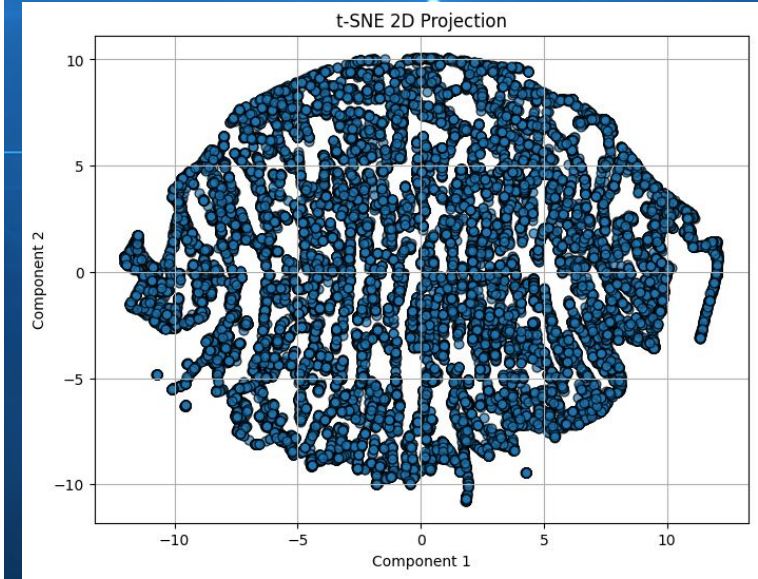
loan_status	count
0	25321
1	7088

target	count
0	25321
1	25321

PCA (linear) and t-SNE (non-linear)

- **PCA**: Dimensionality reduction with principal components, maximum variance.
- **t-SNE**: Using Gaussian and t-distribution, similarity score.





No cluster formation in both.

Supervised Learning: Random Forest

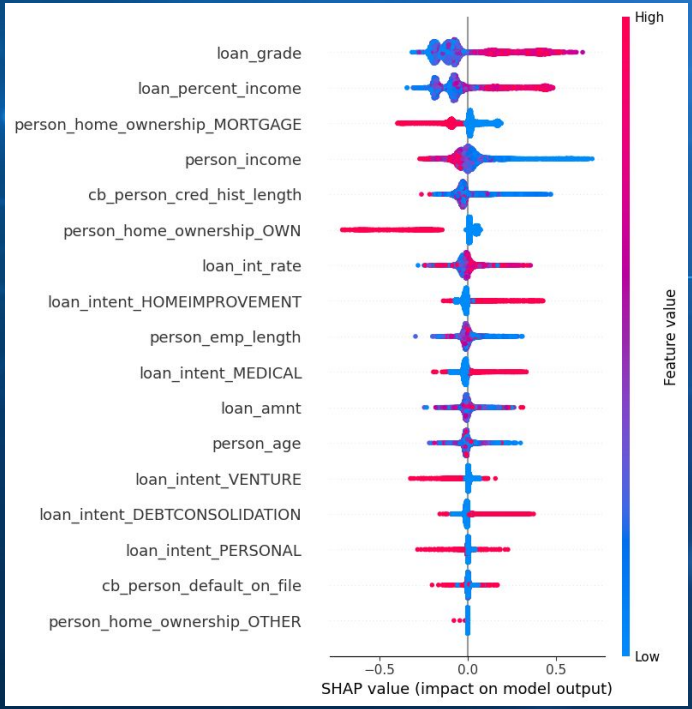
- Random forest classifier
  - Hierarchical tree of boolean grouping

- K-fold: Grid-search integration
  - Test data was 20%
  - Scoring: accuracy
  - Tree depth [VIF]: 26
  - Tree depth [ALL]: 29
  - Folds: 5

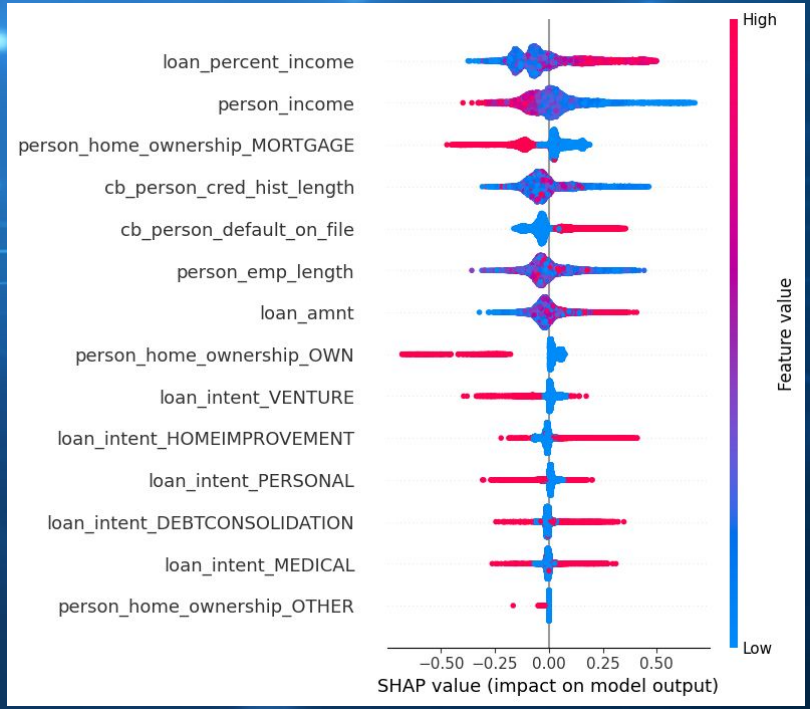
$$accuracy = \frac{\#\ correctly\ classified}{\#total\ number}$$

Random Forest - Results

[ALL] Accuracy: 0.940 F1: 0.940



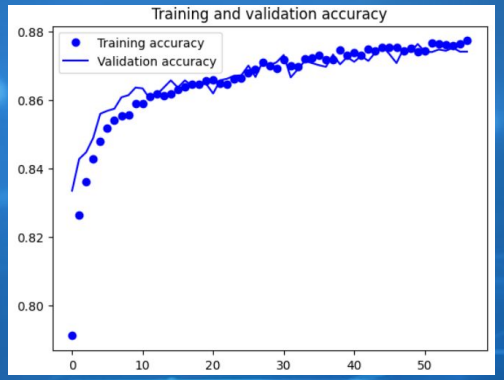
[VIF < 10] Accuracy: 0.899 F1: 0.900

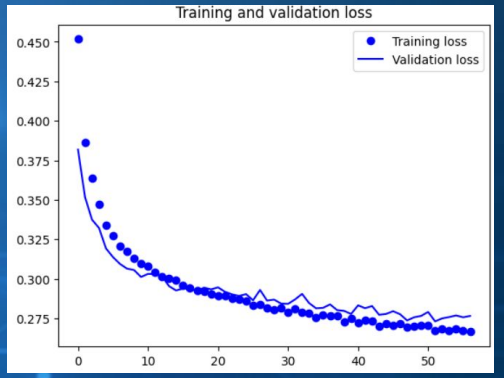


Neural networks: all data

Model 1:
3 hidden layers, leakyReLU + Dropout(0.3)
, EarlyStopping(patience=5)

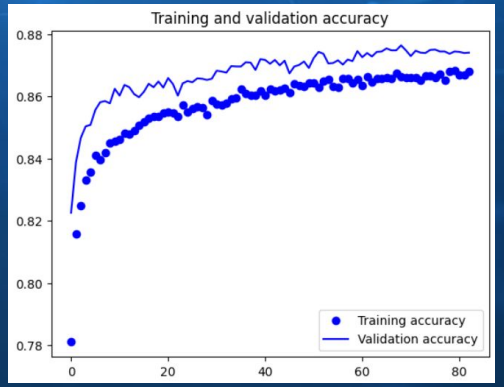
```
Test loss: ~0.270
Test accuracy: ~0.871
```

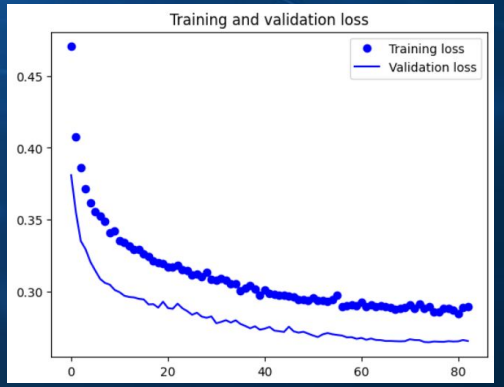




Model 2:
4 hidden layers
, BatchNorm + leakyReLU + Dropout(0.3)
, EarlyStopping(patience=8)
+ ReduceLROnPlateau

```
Test loss: ~0.274
Test accuracy: ~0.869
```

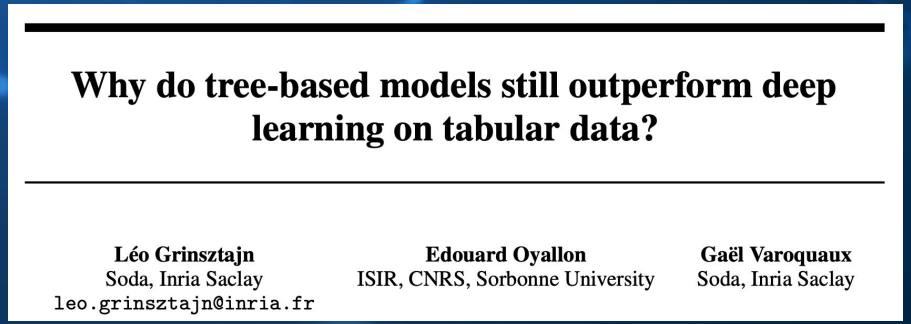




Conclusion

- Neural Network performs slightly worse than Random Forest.
- NN models smooth and continuous pattern. Data shows more irregular, non-continous relationships.
- RF handles these better via recursive splits.
- NN are sensitive to:
  - Feature scaling and orientation
  - Weakly correlated or uninformative features.

- FAIR-principles
  - Random seeds
  - Github



Why do tree-based models still outperform deep learning on tabular data?

Léo Grinsztajn
Soda, Inria Saclay
leo.grinsztajn@inria.fr

Edouard Oyallon
ISIR, CNRS, Sorbonne University

Gaël Varoquaux
Soda, Inria Saclay

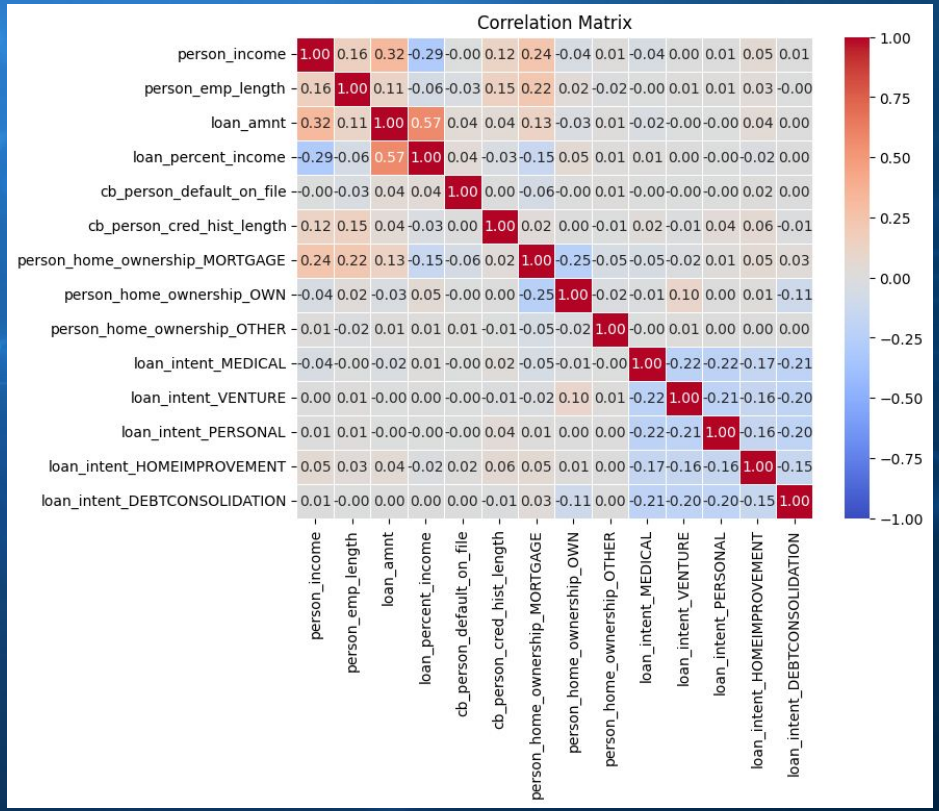
References

1. Gustafsson, A. (n.d.). Exercise 2 – Random forest biome modelling. Lund University.
2. James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An introduction to statistical learning: With applications in Python. Springer.
3. Mentioned paper: https://arxiv.org/abs/2405.01978

VIF filtering and Collinearity

- VIF < 10
- Derived from balancing highly correlated input combinations

- Original data: 17 variables
- Filtered data: 14 variables

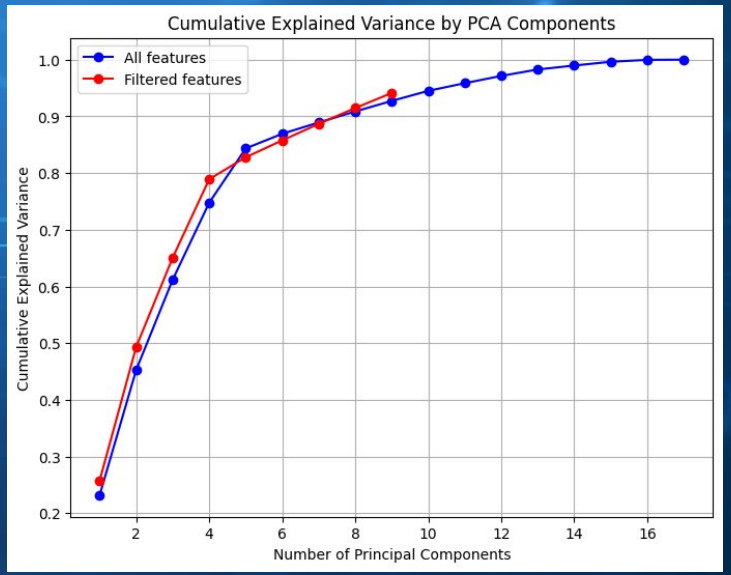


Unsupervised Learning: PCA

- Dimensionality reduction with principal components, capturing the variance

- Considers linear relationships

  - Component 1: ~ 23 % of total variance
  - Component 2: ~ 22 % of total variance
  - Component 3: ~ 16 % of total variance

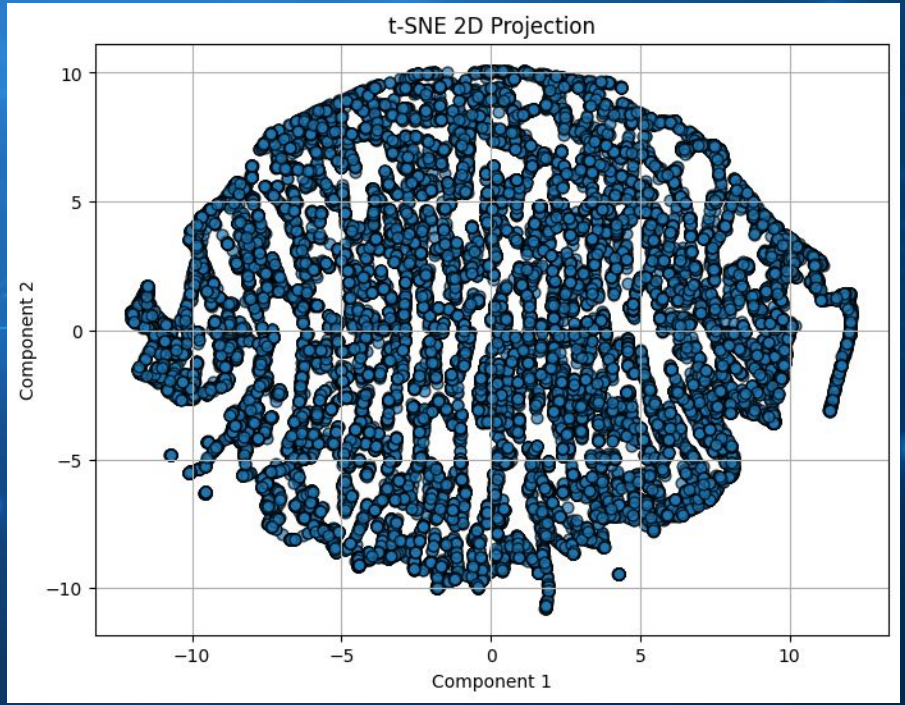


Fails to reduce information into few principal components

→ There is no distinct clusters, poor clustering

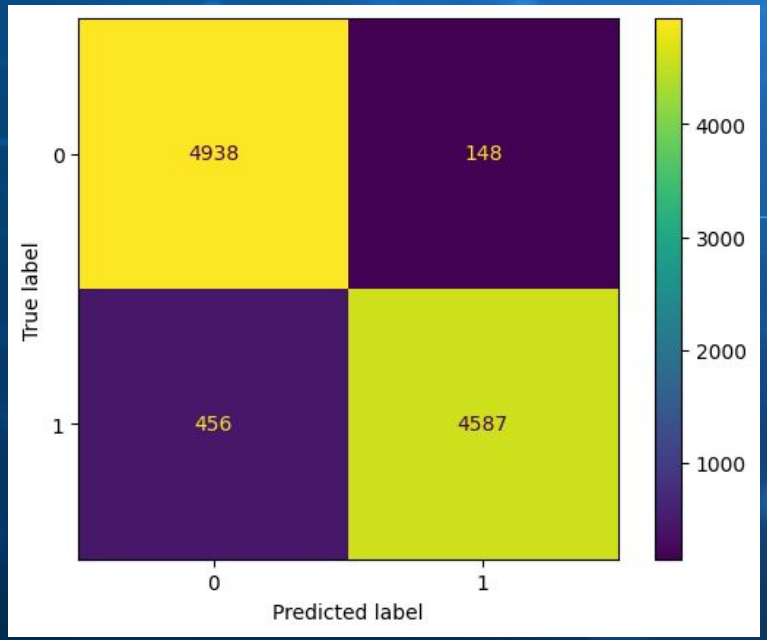
Unsupervised Learning: t-SNE

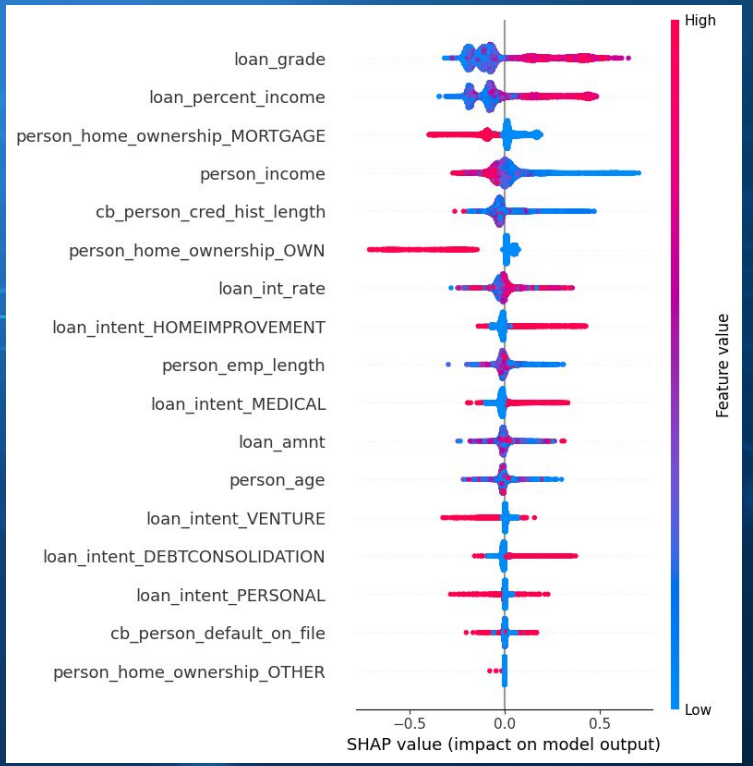
- Considers non-linear relationships
- Using t-distribution, similarity score
- No grouping of data → poor clustering



Random Forest - Results [ALL]

Accuracy: 0.940 F1: 0.940





Random Forest - Results [VIF filter]

Accuracy: 0.899 F1: 0.900