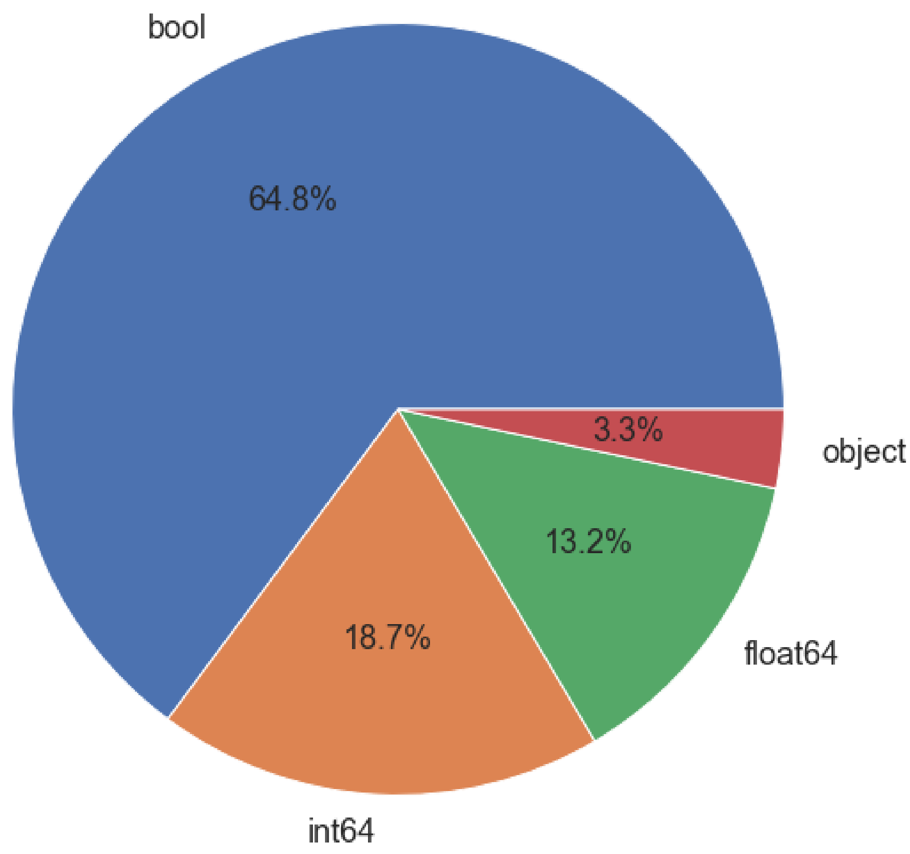
EDA and ML on Prosper Data (ongoing)

D7: Crowdfunding Default/Fraud Detection

Nadi Serhan Aydın, PhD, FRM

Data pre-processing



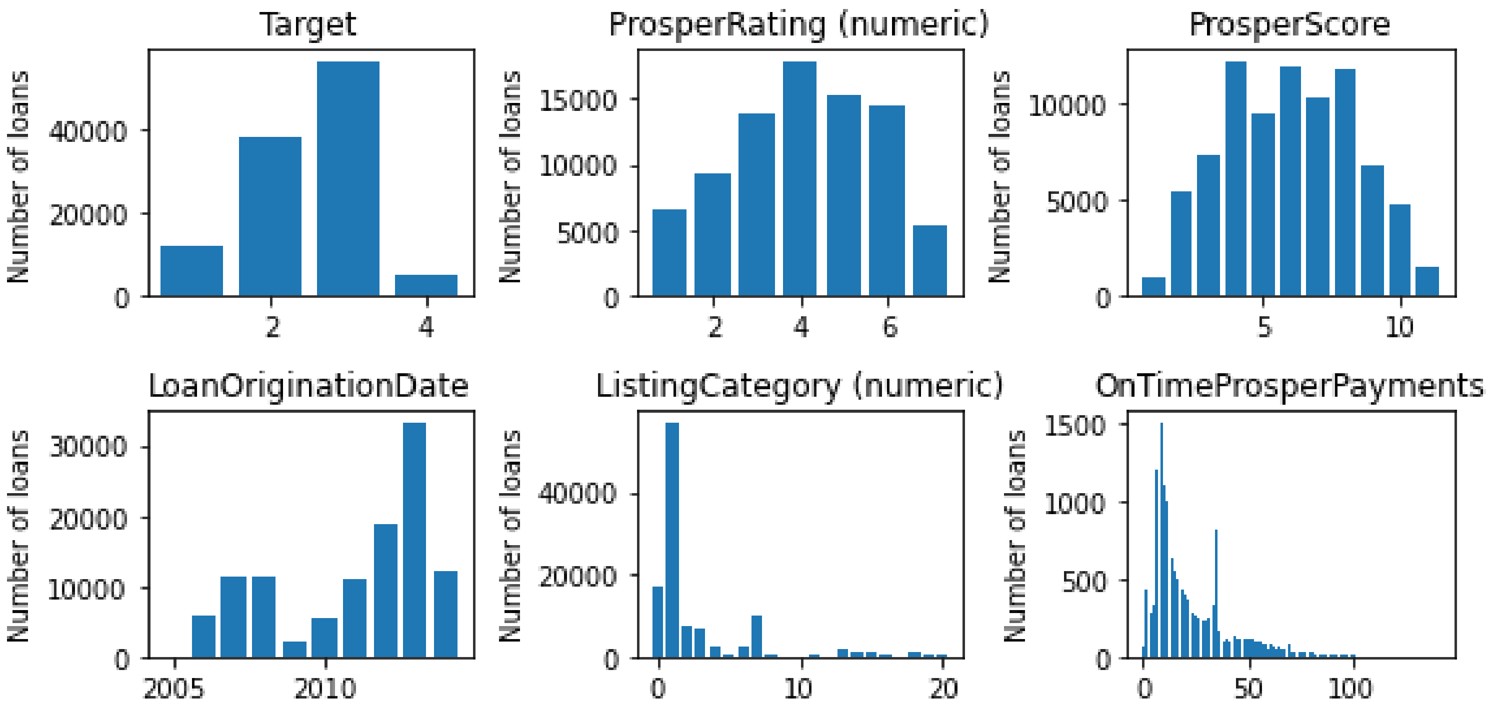
Data pre-processing

- Initial data shape: (113937, 81)
- Target variable: LoanStatus
- Produce % columns out of some absolute value columns
- Remove labels that are not logical or have scarce data: (55084, 91)

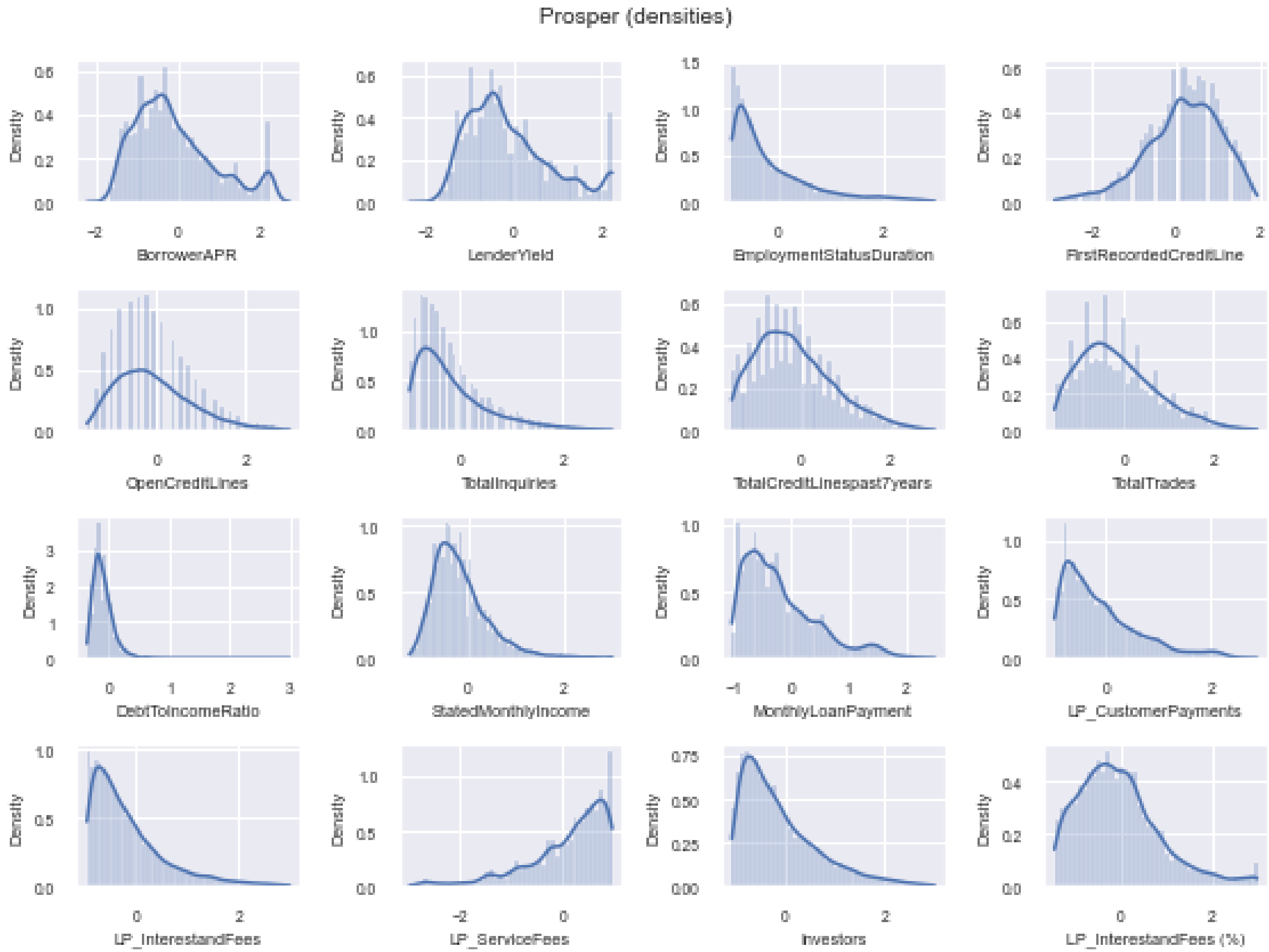
```
list(np.unique(data.LoanStatus))
✓ 0.0s

['Chargedoff', 'Completed', 'Defaulted']
```

- Remove features with >50% of data missing
- Drop remaining rows with at least with one NaN value: (18506, 67)
- Convert Boolean and categorical variables to integers
- Apply standard scaler and remove outliers

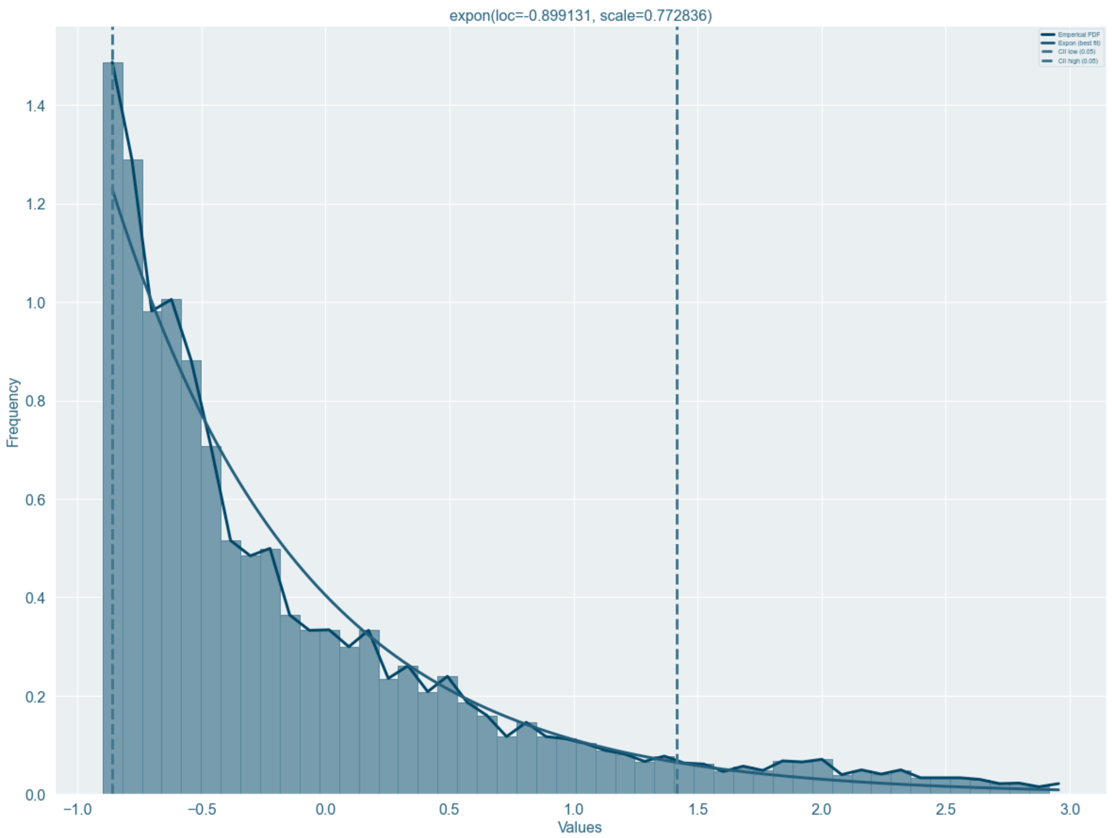
Pre-processed data



Empirical densities of select features



Estimated densities of select features



EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY
FIN AI
Fintech and Artificial Intelligence in Finance

Feature selection

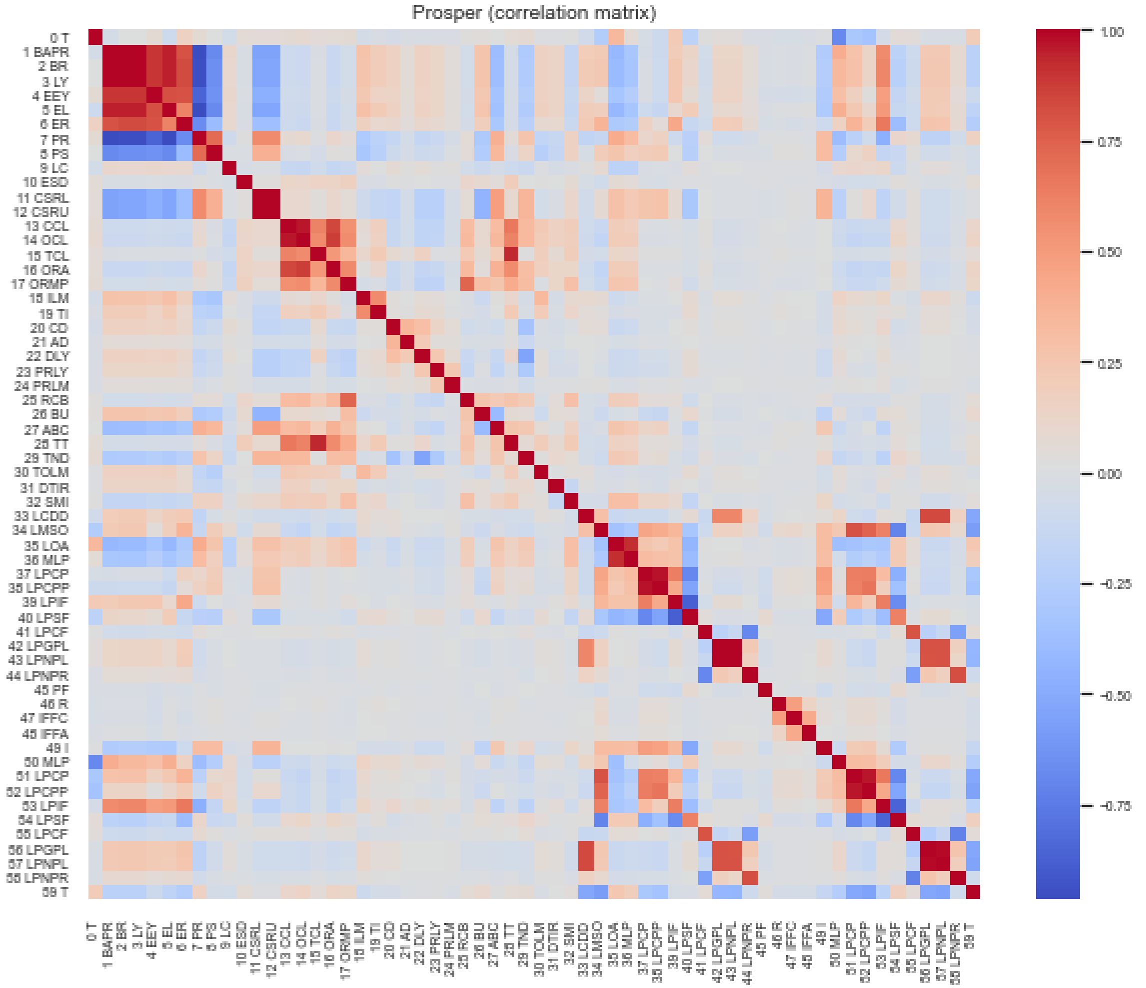
- Feature selection is used when we you know the target variable (Supervised Learning).
- For Unsupervised Learning, there is no exact technique.
- Dimensionality Reduction can be used to reduce the number of features and give us the core set of features which can explain most of the variability in the dataset.
  - The features would be derived from the existing features and might or might not be the same features.
- There are different techniques which are available for doing so:
  - PCA, Linear discriminant analysis, Non-negative Matrix Factorization, Generalized discriminant analysis, etc.

Feature selection

- One feature selection scheme is to select features that correlate strongest to the classification variable. This has been called **maximum-relevance** selection.
  - Many heuristic algorithms can be used, such as the sequential forward, backward, or floating selections.
- On the other hand, features can be selected to be mutually far away from each other while still having high correlation to the classification variable.
  - This scheme, termed as **Minimum Redundancy Maximum Relevance (MRMR)** selection has been found to be more powerful than the maximum relevance selection.
  - Chi2, ANOVA, Kruskal Wallis can also be used.

Raw correlation matrix

- 66 features + LoanStatus

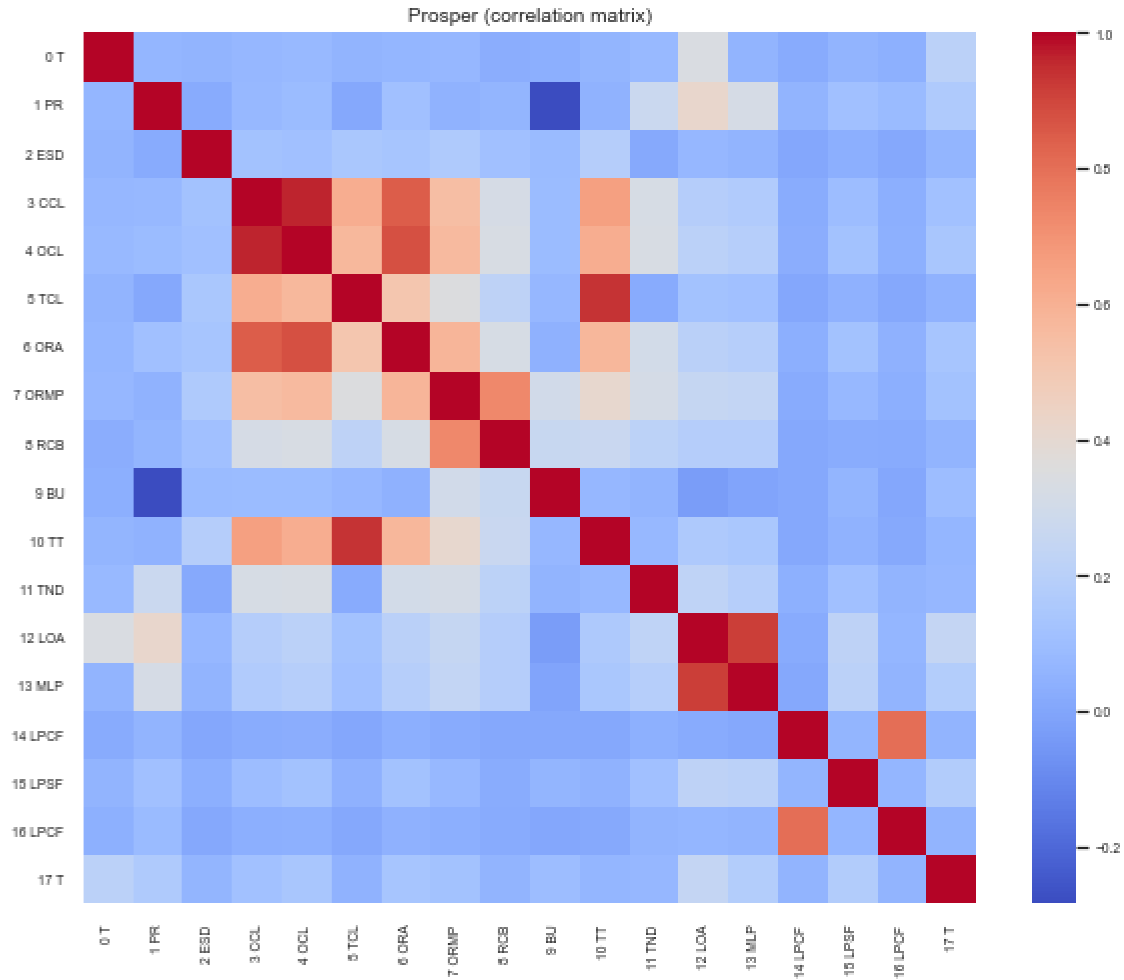


Raw correlation matrix

- Let's keep in mind that correlation is a measure of **linear** relationship
- An ML model can discover **non-linear** relationships as well
- We further **refine the feature matrix** through Maximum Relevance Minimum Redundancy (MRMR)

Refined correlation matrix

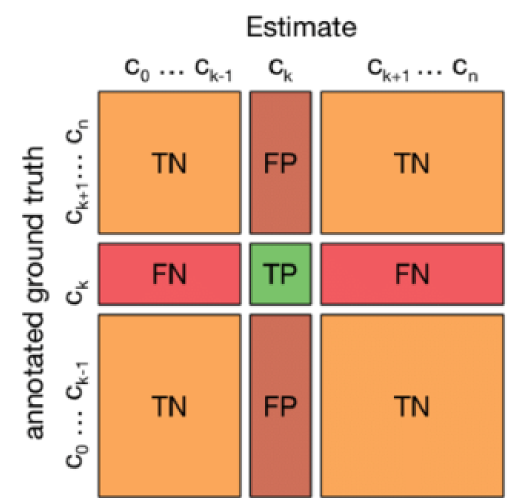
- 12 features + LoanStatus



Performance metrics

- Accuracy: $\frac{TP+TN}{P+N}$
  - Can be magnified by the high number of TN
- Precision: $\frac{TP}{TP+FP} = 1 - \frac{FP}{TP+FP}$
  - Matters when the cost of FP is high (spam emails)
  - Focuses on reducing Type 1 error
  - What % of predicted positives are true positives?
- Recall: $\frac{TP}{TP+FN} = 1 - \frac{FN}{TP+FN}$
  - Matters when the cost of FN is high (fraud activity)
  - Focuses on reducing Type 2 error
  - What % of actual positives are predicted positives?
- F1 score: $2\left(\frac{\frac{TP}{TP+FP}\cdot\frac{TP}{TP+FN}}{\frac{TP}{TP+FP}+\frac{TP}{TP+FN}}\right) = 2\left(\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}\right)$
  - Balances between precision and score
  - Better than accuracy when actual N is too large

		Predicted	
		Negative **(N)** -	Positive **(P)** +
Actual	Negative -	True Negative **(TN)**	**False Positive (FP) Type I Error**
	Positive +	**False Negative (FN) Type II Error**	True Positive **(TP)**



Performance results

- Without re-sampling (80-20 train-test)

	tr_acc	tr_prec	tr_rec	tr_f1	te_acc	te_prec	te_rec	te_f1
logistic	0.6922	0.4792	0.6922	0.5664	0.6956	0.4838	0.6956	0.5707
svc_poly	0.6928	0.7392	0.6928	0.5679	0.6953	0.4838	0.6953	0.5706
svc_rbf	0.6933	0.7874	0.6933	0.5687	0.6956	0.4838	0.6956	0.5707
svc_lin	0.6922	0.4792	0.6922	0.5664	0.6956	0.4838	0.6956	0.5707
lin_svc	0.6922	0.4792	0.6922	0.5664	0.6956	0.4838	0.6956	0.5707
dec tree	0.9990	0.9990	0.9990	0.9990	0.5146	0.5338	0.5146	0.5238
sto_grad	0.6887	0.5433	0.6887	0.5681	0.6953	0.5730	0.6953	0.5742
knn	0.7145	0.6836	0.7145	0.6659	0.6248	0.5379	0.6248	0.5718
grad boost	0.6931	0.6669	0.6931	0.5684	0.6945	0.4836	0.6945	0.5702
mlp	0.6933	0.6310	0.6933	0.5715	0.6945	0.5557	0.6945	0.5727

Performance results

- Random over-sampling (80-20 train-test)

	tr_acc	tr_prec	tr_rec	tr_f1	te_acc	te_prec	te_rec	te_f1
logistic	0.351649	0.351194	0.351649	0.349907	0.353782	0.353841	0.353782	0.352576
svc_poly	0.381446	0.38906	0.381446	0.37459	0.366389	0.3734	0.366389	0.360147
svc_rbf	0.423428	0.423814	0.423428	0.420671	0.392644	0.391552	0.392644	0.389657
svc_lin	0.351779	0.352057	0.351779	0.34381	0.351573	0.352182	0.351573	0.343538
lin_svc	0.352851	0.352613	0.352851	0.350815	0.354822	0.354912	0.354822	0.35328
dec tree	0.999058	0.999059	0.999058	0.999058	0.852742	0.863506	0.852742	0.844879
sto_grad	0.335695	0.334719	0.335695	0.313638	0.337666	0.335249	0.337666	0.3171
knn	0.801885	0.80913	0.801885	0.790084	0.686249	0.676691	0.686249	0.666203
grad boost	0.768383	0.766933	0.768383	0.766456	0.658305	0.65388	0.658305	0.654045
mlp	0.473404	0.470109	0.473404	0.462457	0.442943	0.436104	0.442943	0.431352

Research directions

I. Pre-/post-fraud-policy change comparison
   - Not too much related to ML
II. Try labeling existing data
   - Fraud is a legal issue (model-based labelling is not a choice)
   - Fraud indicators (even quantitative ones) hardly apply to historical data ()
III. Work on target variables at hand just to showcase that data has some explanatory power
IV. …