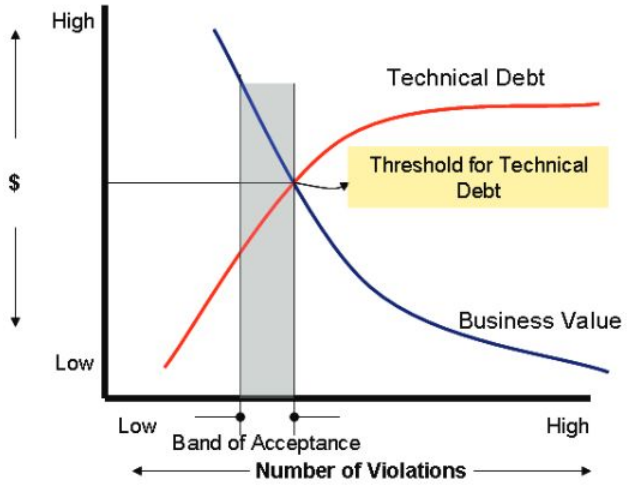
TAED2 Software Analytics Project

The Data Miners

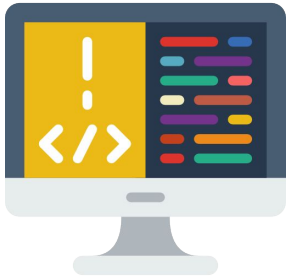
Background

- Technical Debt dataset
- SonarQube and Jira



Business and Project Goal

Code issues: Sonar measures impact.



Help developers: Detecting patterns.

Success criteria and benefits

Model: Regression, Neural Networks.

Results: Patterns, correlations.

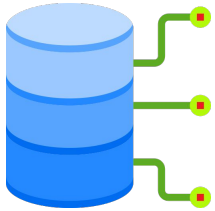
Benefits:

- Efficiency.
- Save resources.

Project Plan

Business understanding

Data understanding



Data preparation

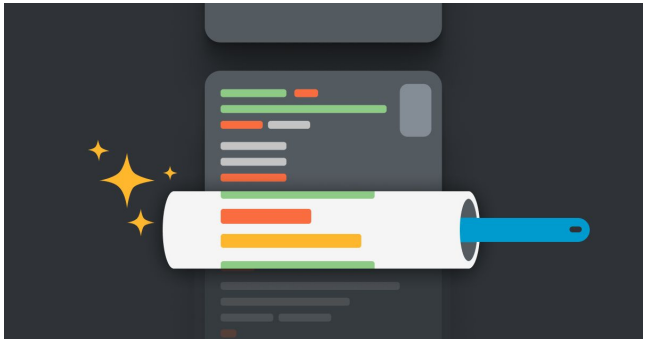
Modeling

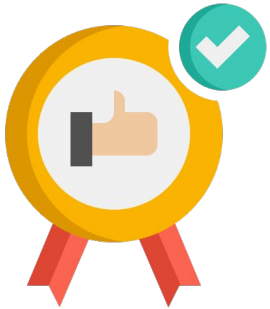
Evaluation

Deployment

Good engineering practices

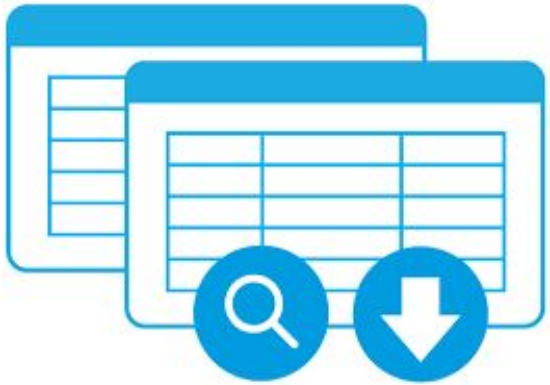
- File structure and replication package (“*Cookiecutter template*”).
- Using linters for code quality (*mllint*).
- Sharing status and outcomes of experiments within the team.





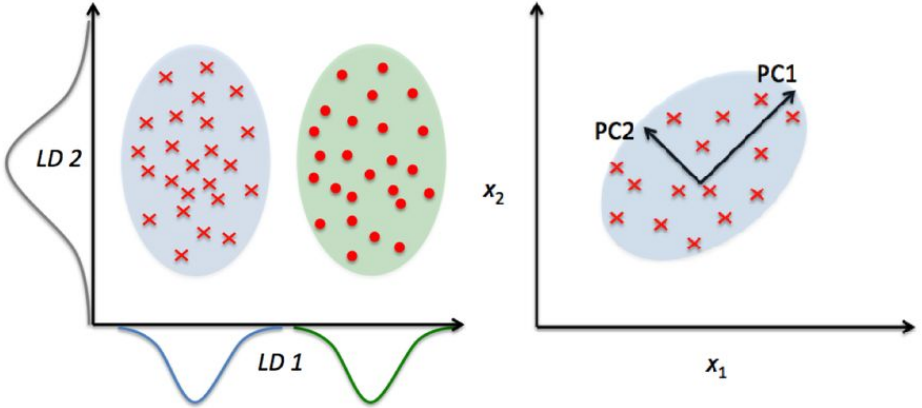
Data Selection

- Sonar Measures
  - Complexity
  - Sqale debt
- Sonar Issues
  - Type
  - Fixing time



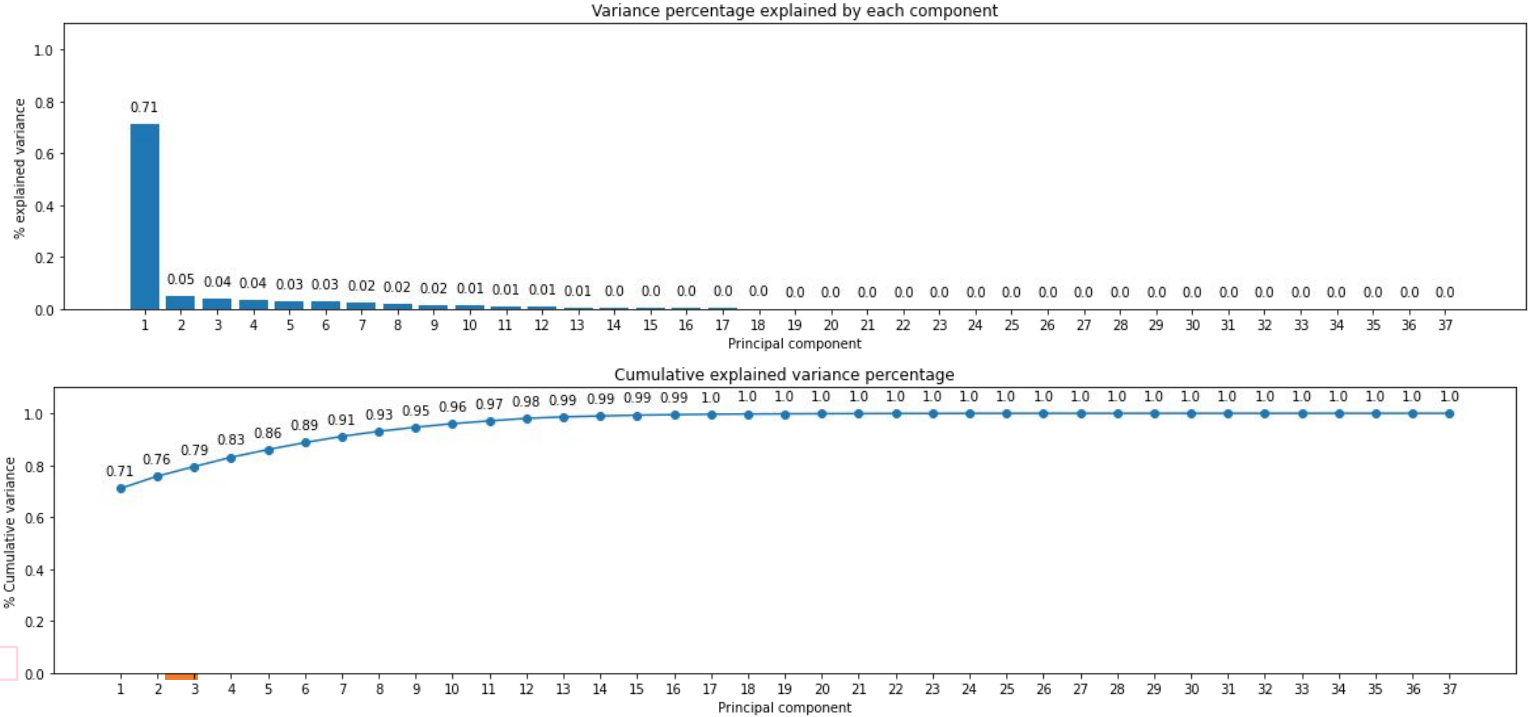
Modelling

- Principal component analysis (PCA)
  - Reduce dimensionality
  - Easier computation
  - See correlation in the data



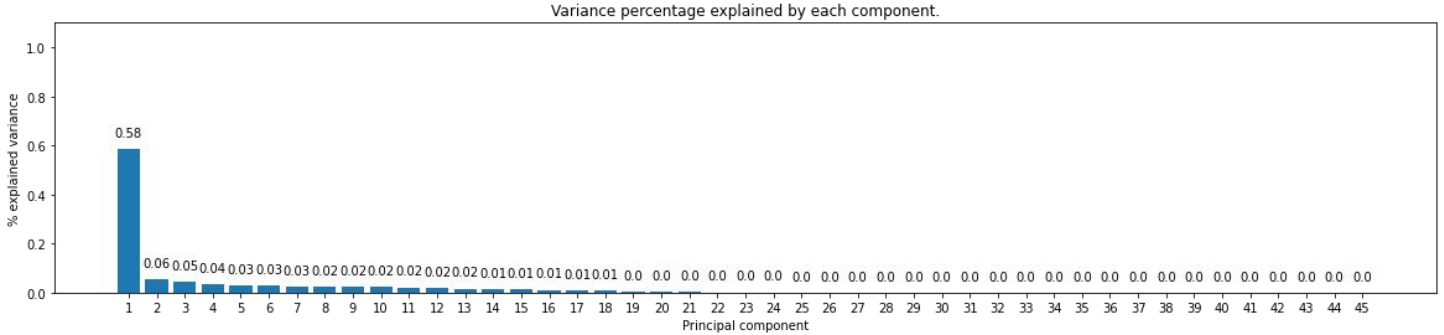
PCA

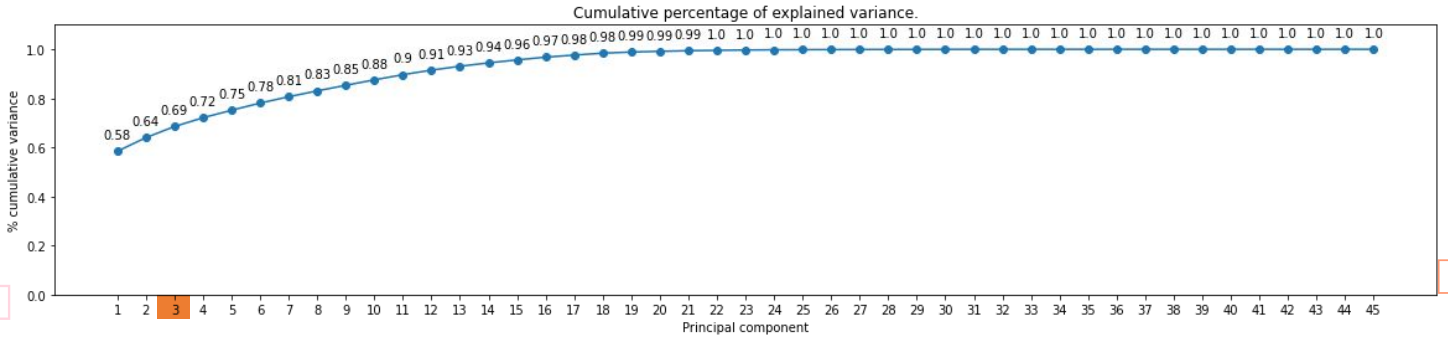
- Using only numerical variables



PCA

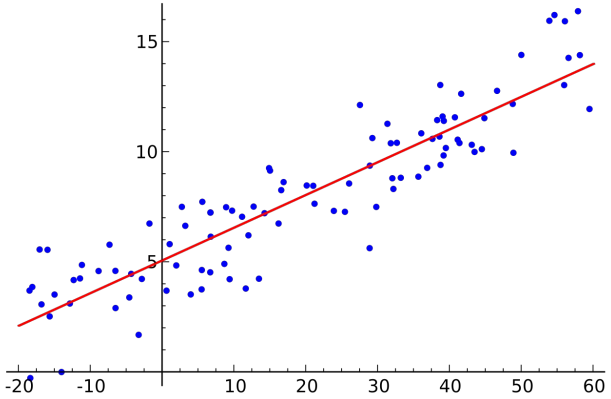
- Adding categorical variables

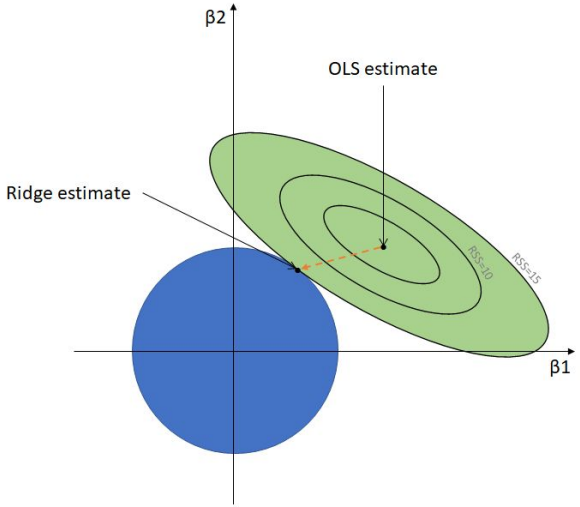




Regression

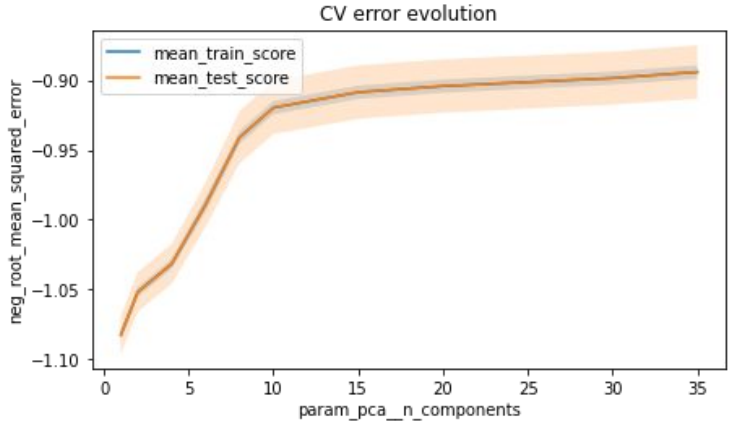
- Linear regression
- Principal components regression
- Grid search with cross validation
- Ridge regression

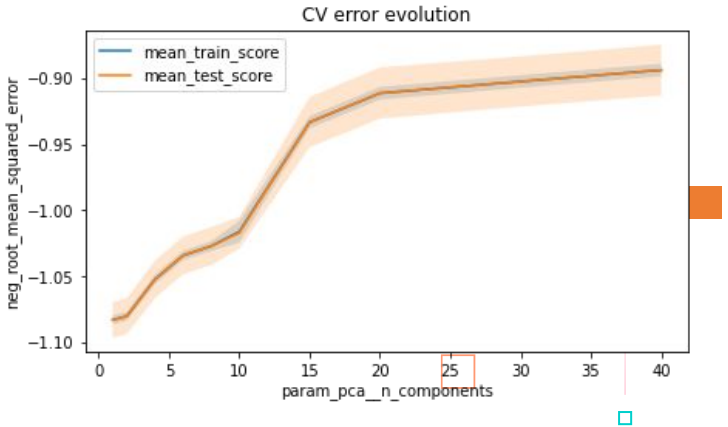




Evaluation

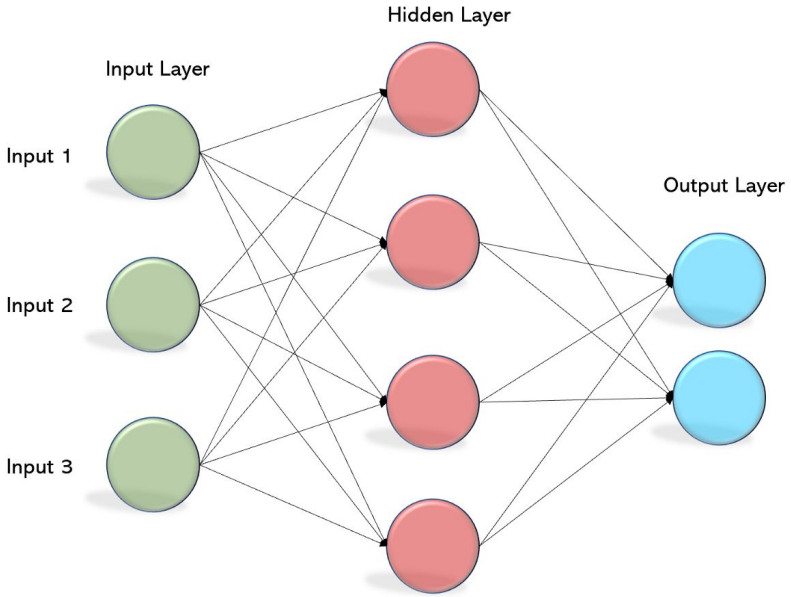
Regression	Error	Type of error
Linear regression	0.890	rMSE
PCA numerical data	1.011	rMSE
PCA one hot encoding	0.874	rMSE
Ridge regression (α = 1)	0.352	MAE





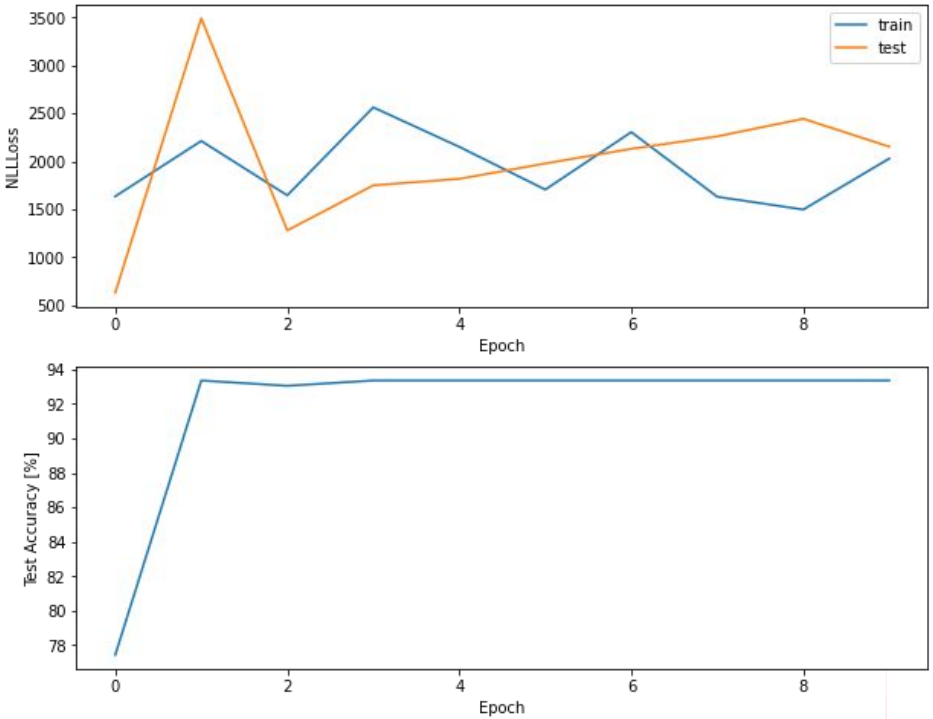
Multilayer perceptron

- Supervised learning
  - Hyperparameters
- Classification/Prediction
  - Sonar issue type
  - Any categorical variable



Evaluation

- Big batch size -> High loss
- Small batch size -> Overfitting
- 93% accuracy



Conclusions

Model

- Ridge regression with one hot encoded data
  - Error: 0.352% with $\alpha = 1$
- Successfully predicts the technical debt ratio

Future data mining

- Our work is replicable
  - Changing the variables in our code