Learned from literature:

* Consistently solid performance from Logistic regression (Yu et al., Alp et al., Yang and Zhang), couple with the fact that it does not rely on hyperparameters
* Avoided pitfall of upward biased performance estimates using cross-validation after oversampling and included the oversampling only for the training folds. (Blagus and Lusa)
* Inclusion of XGBoost as a candidate (Yang and Zhang)

Outline:

1. Literature Review
   1. Paper 1
   2. Paper 2
   3. Paper 3
   4. Paper 4
   5. Paper 5
   6. Summary and how it affected the work (above^)
2. Report on data-mining
   1. Business understanding & Data understanding
      1. The understanding of the client and our assumptions:
         1. Dataset of historical credit applications
         2. In current system, there are risk averse credit processors working for the client who are manually assessing applications based on their experience.
      2. Some points from market research (Discuss with others)
      3. Dataset exploration
         1. Imbalanced class problem
            1. Add plot on imbalance
         2. Few missing values
         3. Types of information included (Just briefly: Pieces of socio-economic information, credit card history, History with Universal credit)
            1. Refer to appendix with univariate plots of all variables
   2. Data preparation
      1. Dropping missing data due to its low volume and in this case hardly replacable nature.
      2. Dataset definitions
         1. Sparse: PY =-1 and PY=-2 simplified to PY=0
         2. Rich: separate out PY into paidfull and notransaction
      3. Correcting class imbalance
         1. Random Oversampling
         2. Random Undersampling
         3. SMOTE
      4. Data standardisation (scaling to normal dist)
         1. contract the scale while preserving high and low values
         2. We do not drop outliers, but preserve them due to the information contained in them
      5. Train – test split + Cross\_validation
         1. 80% training sample
            1. A lot of data, can afford to use as much as 80% of the data.
            2. 5-fold cross-validation to choose dataset, algorithm and hyperparameters
         2. 20% test sample
            1. Hold-out sample until final evaluation
   3. Modeling
      1. Baseline comparison using logistic regression
         1. Logistic regression
            1. information (reference to Elements of statistical learning textbook)
            2. link to literature
            3. no hyperparameters
            4. Given its consistent track record across the literature and its non-reliance on hyperparameters, the logistic regression is used to evaluate, which dataset to use for modeling
         2. Graph about comparison of baseline results
         3. Conclude we use rich\_over
      2. Model comparison
         1. Explain why these models
            1. The key common features of the algorithms compared below include scalability and speed. Although other classification algorithms may also do well, their computational costs tend to scale significantly with observations and variables. As heuristics do not allow to drop any variables in this case, only such methods are explored that scale efficiently and are not prone to overfit when including a high number of features.
            2. Also these are used throughout literature with good results
         2. Penalised logistic regression (reference to Elements of statistical learning textbook)
            1. Variable selection
         3. Decision tree (reference to Elements of statistical learning)
         4. Gradient boosting machine (reference to same book)
         5. XGBoost (reference to Xgboost documentation)
            1. refer back to literature
         6. Results plot, realise that only minor differences
      3. Hyperparameter tuning
         1. Briefly explain what kind of hyperparameters there are and what their general role is. Explain that XGBoost and in general gradient boosting models benefit well from hyperparameter tuning
   4. Evaluation
      1. Modelling evaluation based on test-sample performance, show increasing performance throughout the selection process
      2. Business evaluation through calculation of profit in three scenarios:
         1. 5 input assumptions.
         2. Manual scenario: Client does not adapt our solution
            1. Baseline
         3. Semi-automated scenario: Client uses the model to filter out High-risk and low-risk applications which are automatically rejected or accepted respectively. The rest of the applications are assessed by humans.
            1. High, medium and low risk scenarios with two thresholds
            2. Key assumption: Credit processors work differently and independently from the way model predicts
         4. Automated scenario: Model automatically makes decisions on whether applications are accepted or not
            1. Binary classification system based on optimised threshold.
   5. Conclusion: Recommendation & Deployment strategy
      1. Recommend to implement Semi-automated scenario
      2. Few sentences about what is required for deployment.
         1. Automated input system
         2. Data-format consistency
         3. Processing environment to generate predictions
         4. System to convey medium risk cases to credit processors.
3. References