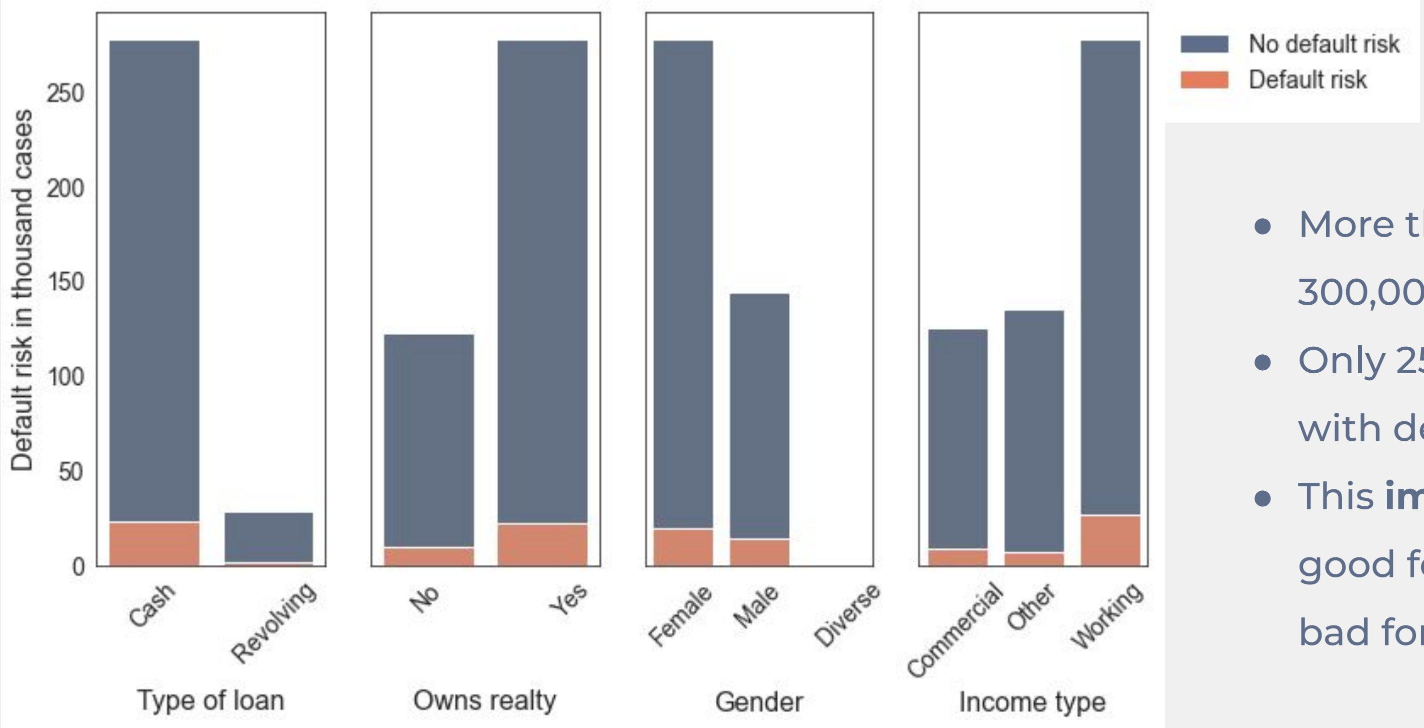
HOME CREDIT

Your life more affordable

GIVING CREDITS TO PEOPLE WHO DON'T GET TRADITIONAL CREDIT

- Give credits to people
- Predict if a person is credit worthy
- Lower default risk by identifying patterns in historical data

FIRST LOOK AT THE DATA



- More than 300,000 customers
- Only 25,000 (8 %) with default risk
- This **imbalance** good for business, bad for prediction

REQUIREMENTS

RUNTIME

Not constrained

FEATURE SELECTION

Selection of best features
from 122 columns

MODELTYPE

Default or no default
> Binary classification

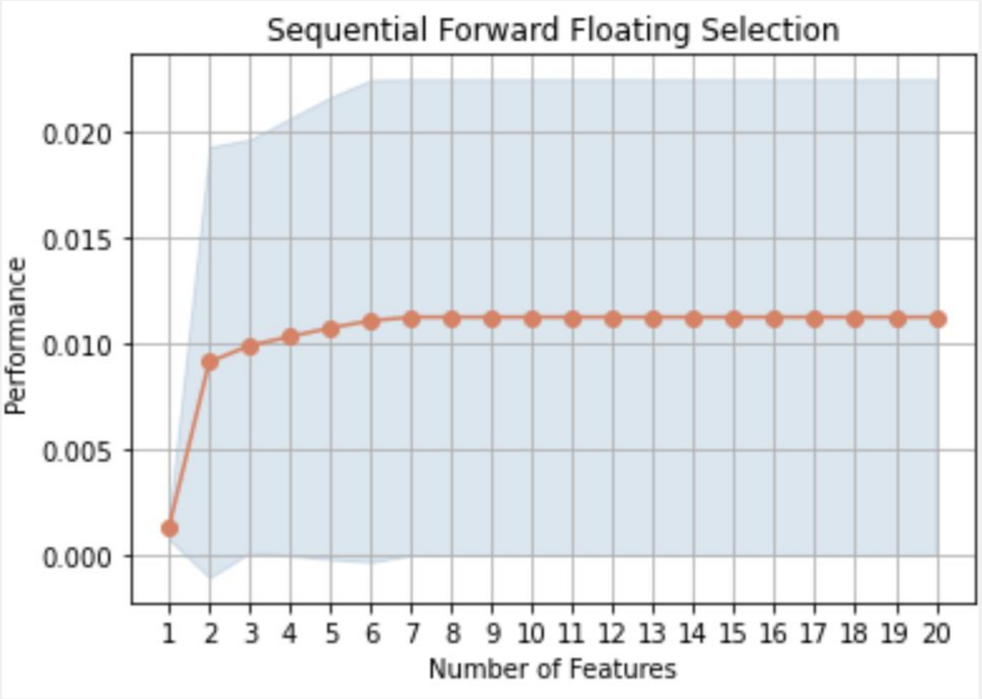
METRICS

Minimize undetected
defaults and false rejection

BALANCING

Resampling train data to
overcome imbalance
and stratify test data

FEATURE SELECTION



- 122 columns in original dataset
  - Age of clients car
  - Days since clients last phone-number change
  - Wall material of clients home
- 178 features in cleaned dataset
- Selection of 20 best features
  - Clients age
  - Income
  - External score
  - Car owner
  - Realty owner

MODEL AND EVALUATION-METRICS SELECTION

Models

Given the loan application data, we don’t have to predict if the applicant is going repay or not in seconds. We can have couple of minutes to predict. Keeping this is mind we consider some Ensemble models like Random Forest and XGboost along with couple of basic models

Evaluation metrics

For an imbalanced data and since desired output is a probability of people defaulting on loan, ROC AUC seems to be better metrics, we also compare recall score (True Positive Rate) as well as accuracy along with ROC AUC

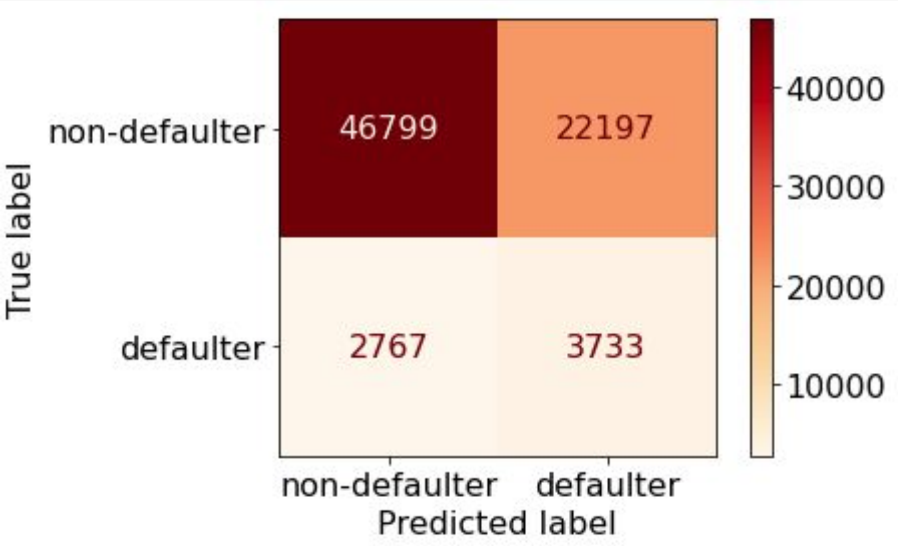
Hyperparameter tuning

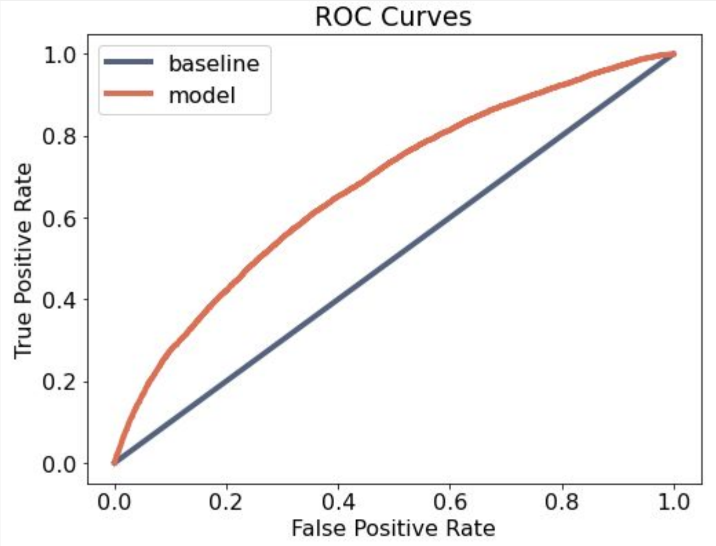
For hyperparameter tuning we can use RandomSearchCV on 2 of our ensemble models and adapt to improve the prediction models.

MODEL OVERVIEW

	HYPER PARAMETER	ROC AUC	RECALL	ACCURACY	TIME
LOGISTIC REGRESSION	None	62 %	57 %	67 %	0.23 s
DECISION TREE CLASSIFIER	None	54 %	53 %	54 %	0.16 s
RANDOM FOREST CLASSIFIER	n_estimators = 196, min_samples_split = 2, max_leaf_nodes = 49, max_depth = 17, bootstrap = True, max_features = 'auto', min_weight_fraction_leaf = 0.1	62 %	54 %	68 %	0.6 s
XGB CLASSIFIER	n_estimators = 200, gamma = 100, learning_rate = 0.01, max_depth = 12, booster = 'gbtree', scale_pos_weight = 1.5, objective = 'binary:logistic'	59 %	83 %	39 %	15.6 s

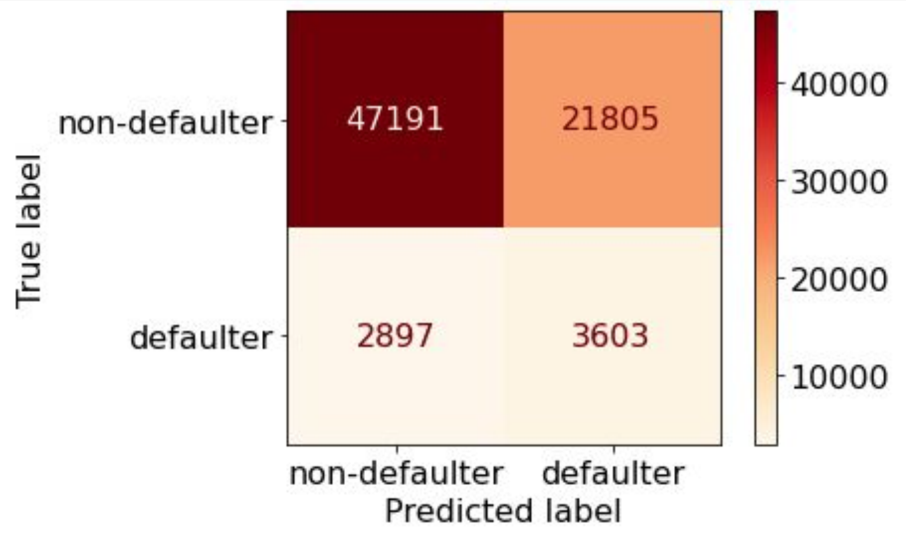
OPTIMIZED LOGISTIC REGRESSION

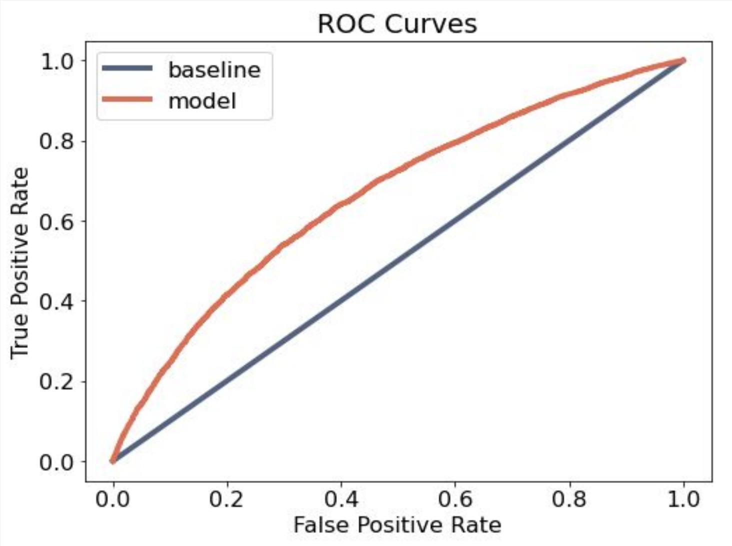




- Simple model
- Fast training and prediction

OPTIMIZED RANDOM FOREST

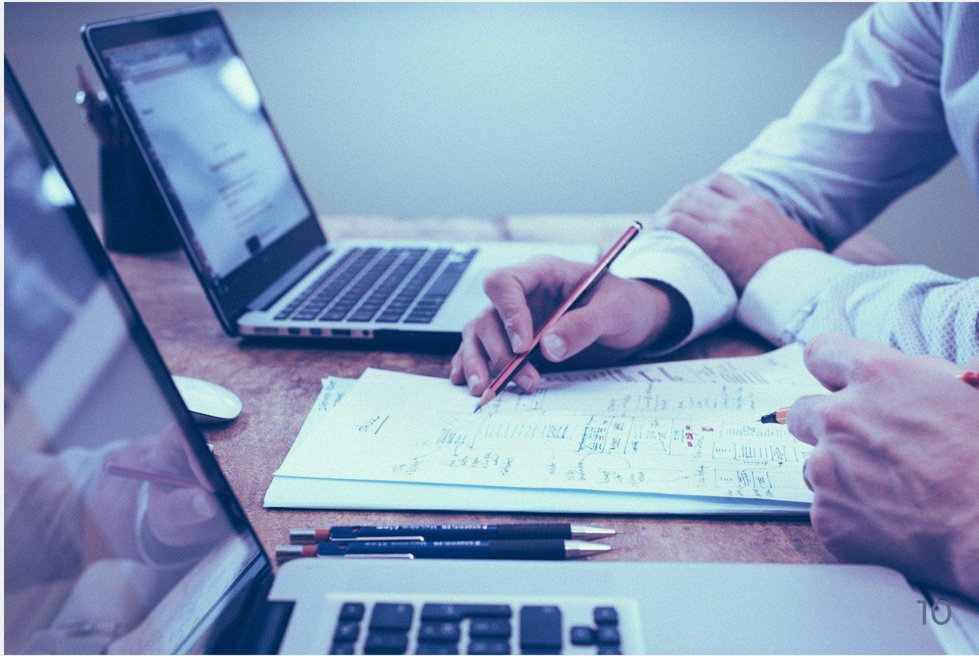




- More good credit (true negative) and less false rejections (false positive)
- Slower than Logistic Regression, but no time restriction

LIMITATIONS AND FUTURE WORK

- Limitations of prediction
  - Missing out on good clients
    > less profit
- Future work
  - Include interest rate and credit term
  - Feature engineering
    (e.g. ratio between income and credit amount)



RECOMMENDATIONS

FEATURE IMPORTANCE

1. Score from external source 2
2. Clients age
3. Owning a car

DATA RECORDING

- \- Housing information etc.
- \+ Interest rate and credit term

CLIENT SCREENING

Reducing administrational cost and increase client base

OUR TEAM

CHANDRA

Modelling Expert

ANDREAS

Number Crusher

Visit our website!

slidesgo