# Predicting default on credit card debt

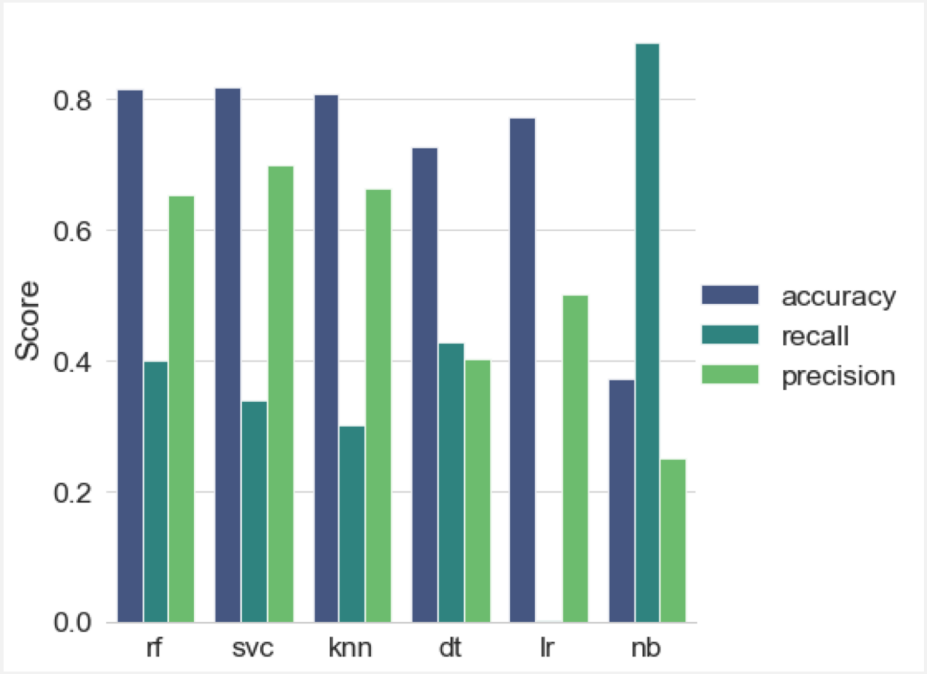
Jason Pizzollo

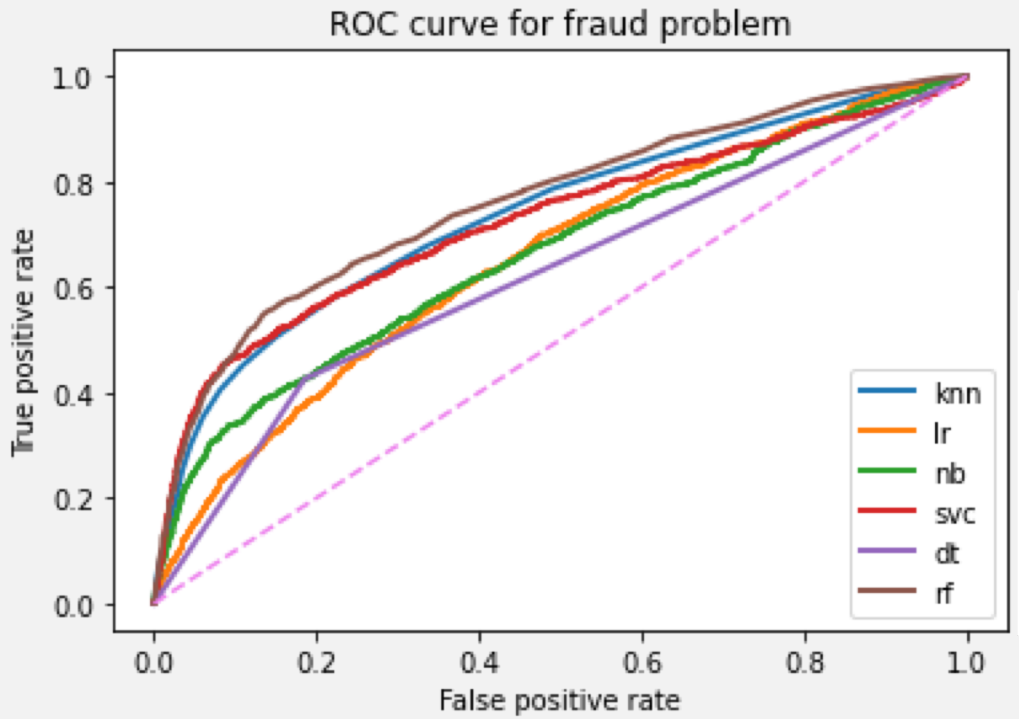
October 28, 2020

# UCI default of credit card clients Data Set

| | |
|---|---|
| Data: | 30,000 customers |
| Features: | 6 months credit card bill |
| | 6 months payment history |
| | Age, Gender, Education Level, Marital Status |
| Label: | Yes/ No (default) |

# Random forest is the best of 6 simple models

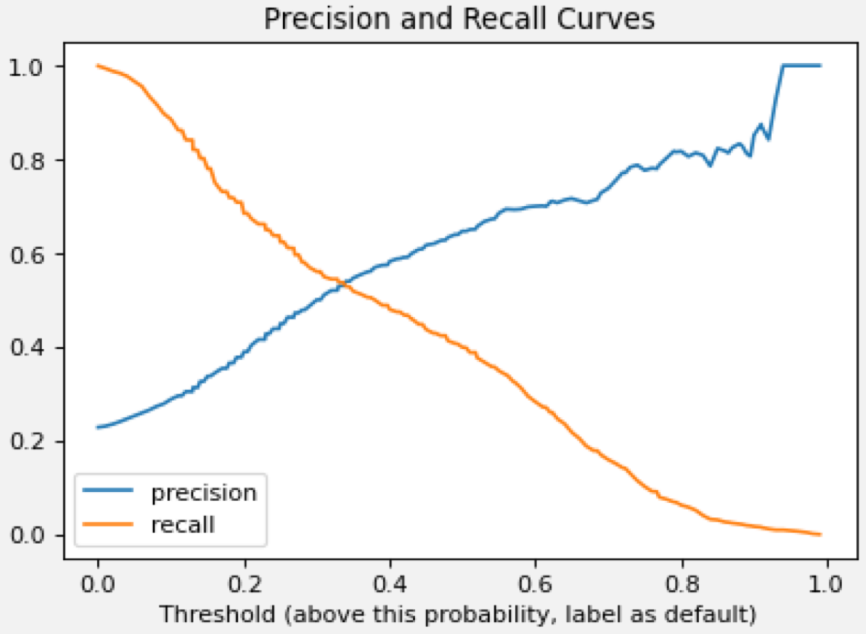



| Model | ROC AUC |
|---|---|
| KNN | 0.74 |
| LR | 0.65 |
| NB | 0.66 |
| SVC | 0.72 |
| DT | 0.62 |
| RF | 0.76 |

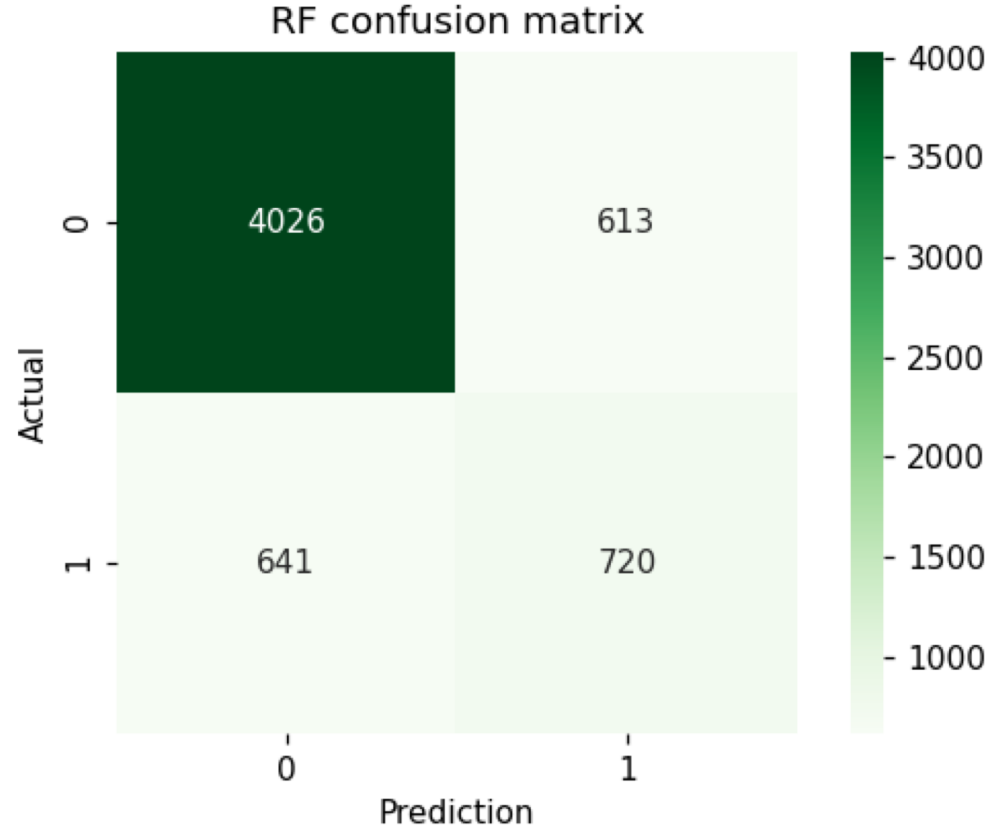
# Optimization

- Marginal accuracy improvement
  - Feature engineering (Debt = Bill - Payment)
  - Balance classes
- Increase default predictions
  - Adjust prediction threshold

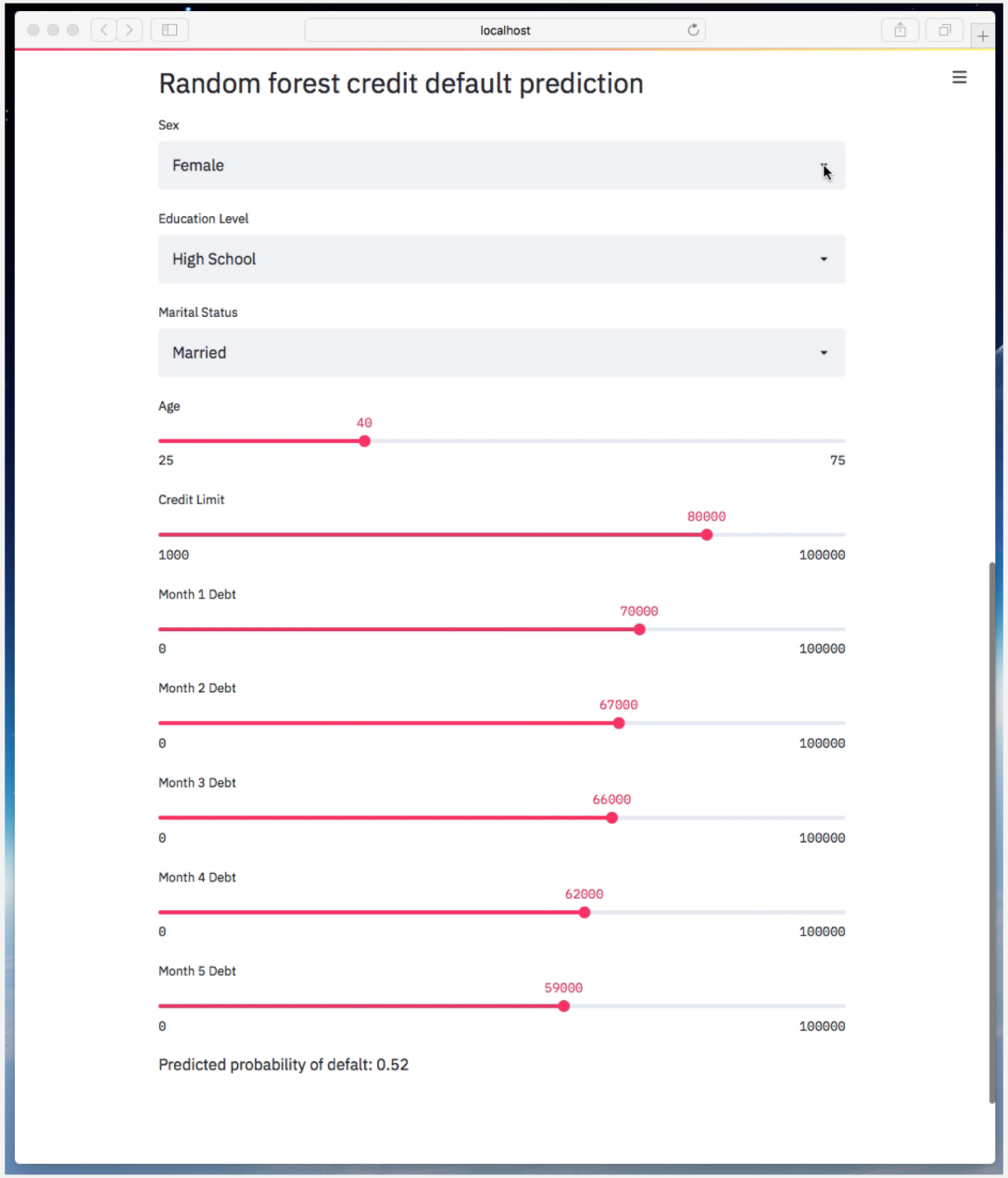



# Final Random Forest Model

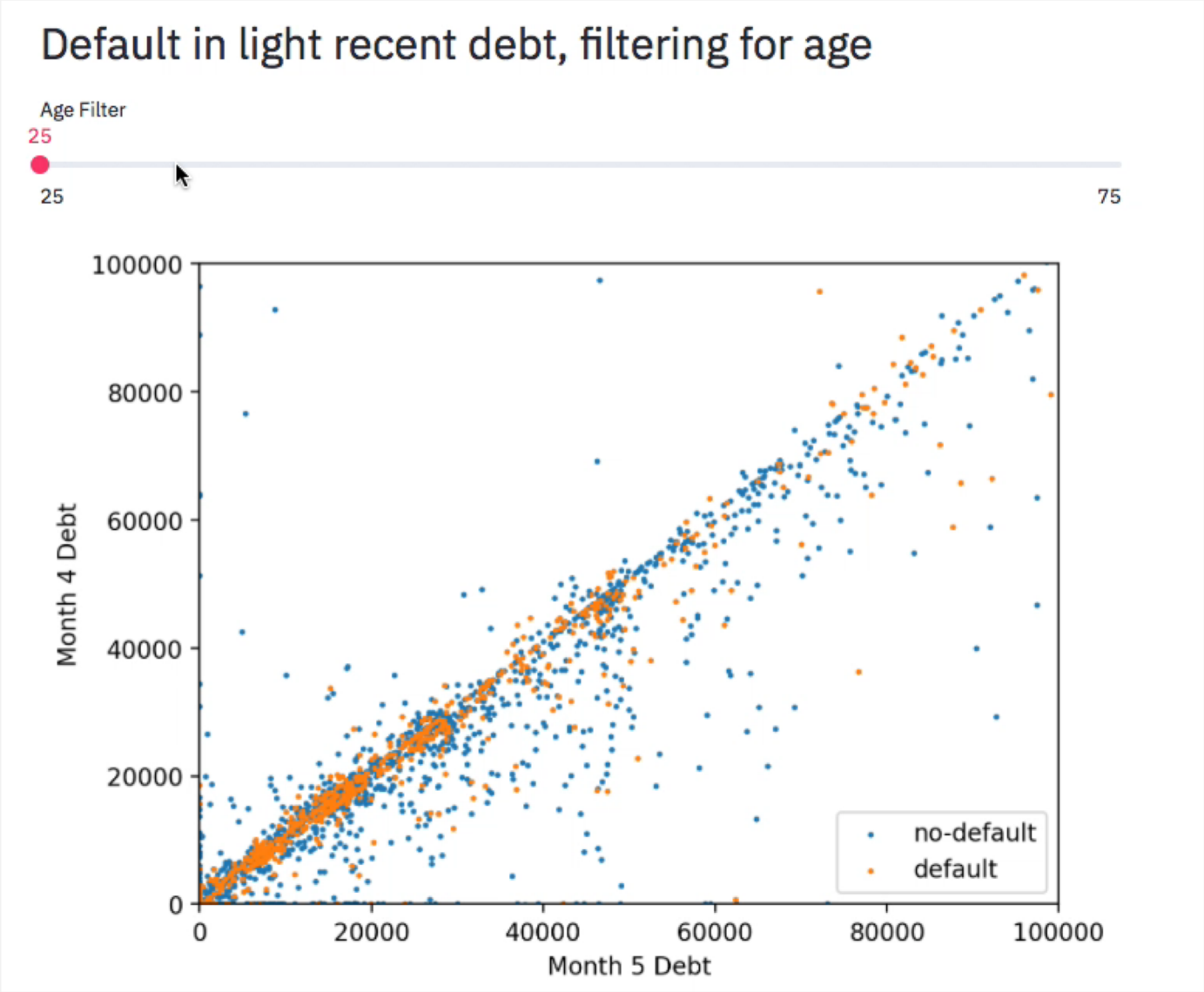
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.86 | 0.87 | 0.87 | 4639 |
| 1 | 0.54 | 0.53 | 0.53 | 1361 |
| accuracy | | | 0.79 | 6000 |
| macro avg | 0.70 | 0.70 | 0.70 | 6000 |
| weighted avg | 0.79 | 0.79 | 0.79 | 6000 |



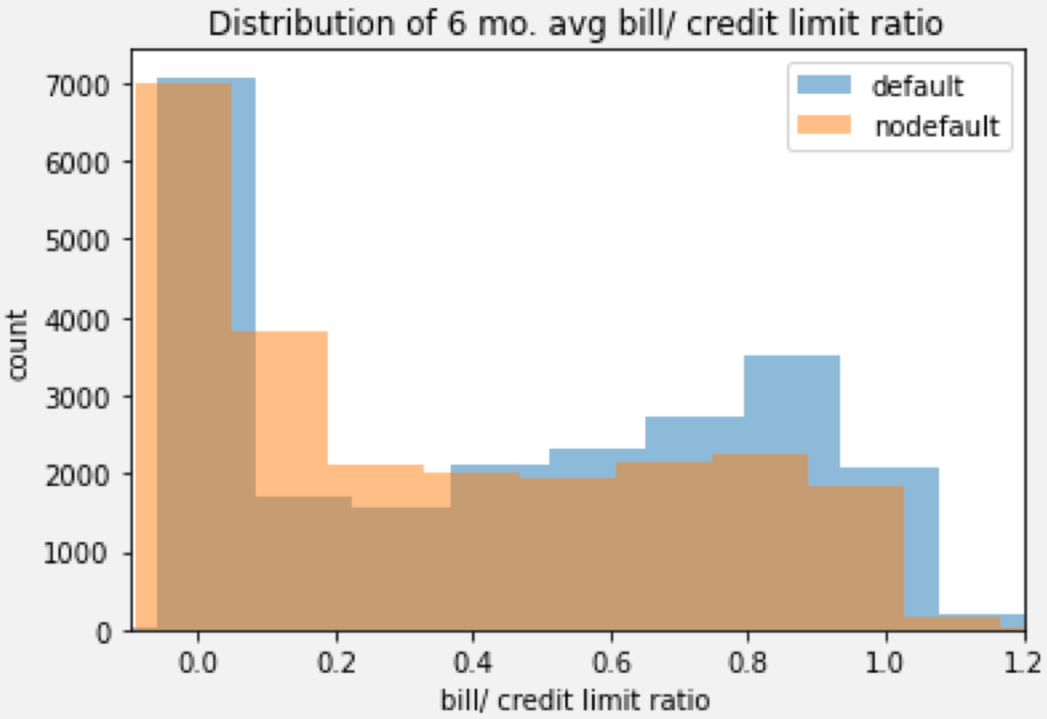
# Prediction Tool



# Visualizing Default ~ debt + age



# How to predict default for customers with low debt?




- Additional data would help:
  - Longer credit history
  - Credit score
  - Employment status
  - Income

# Conclusions

- Noisy data, but…
  - Good accuracy (~0.81) out of the box
  - Edge cases make predictions difficult

# Next Steps

- Making a better model:
  - Ensembling/ boosting/ stacking
  - Additional feature engineering (maybe)