



# Extending a lending hand

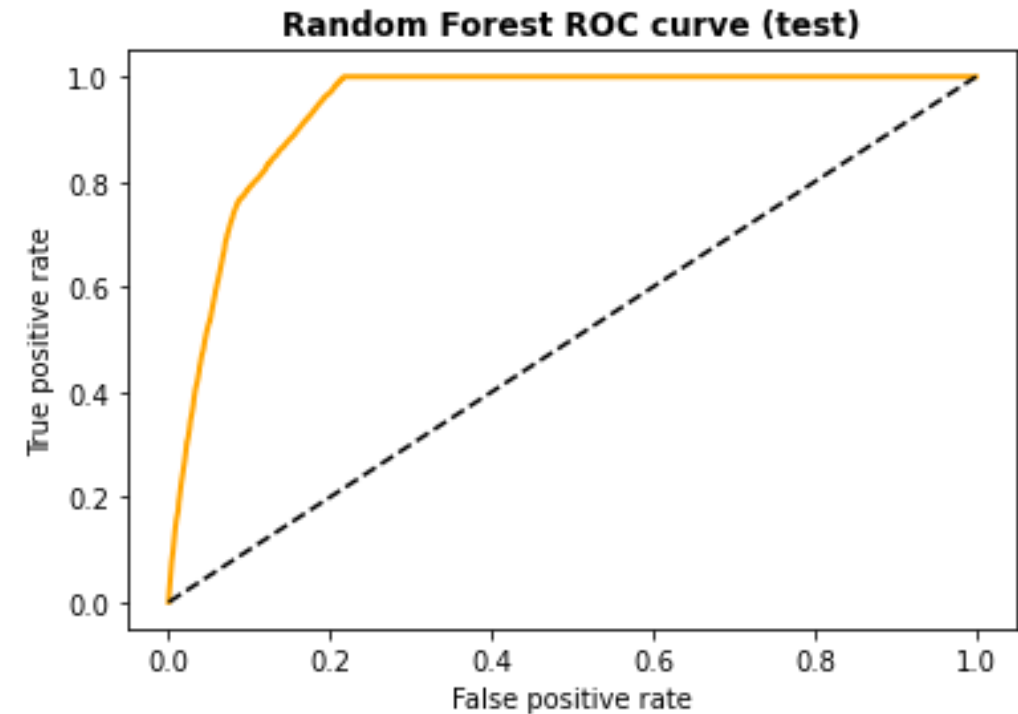
Analyzing consumer credit data to provide financial help to consumers in need

**Sam Reiff**

Sept 2021

# Executive summary

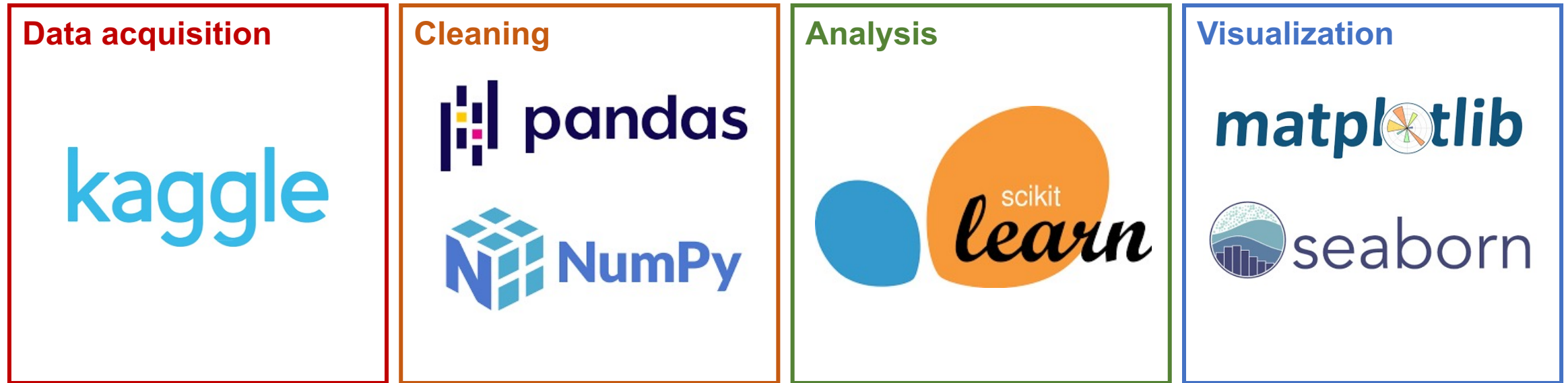
- **Motivation:** Identify default-prone consumers for debt relief outreach
- **Objective:** Establish an effective classification algorithm with acceptable recall/F-beta
- **Conclusion:** Random Forest produces results superior to LR, in-line with KNN with better scalability
  - **Target:** default / no-default
  - Recall on test: 76%
  - F-beta on test: 71%
  - ROC on test: 94%



**Random Forest has reasonable performance metrics and can scale with the business need**

# Methodology

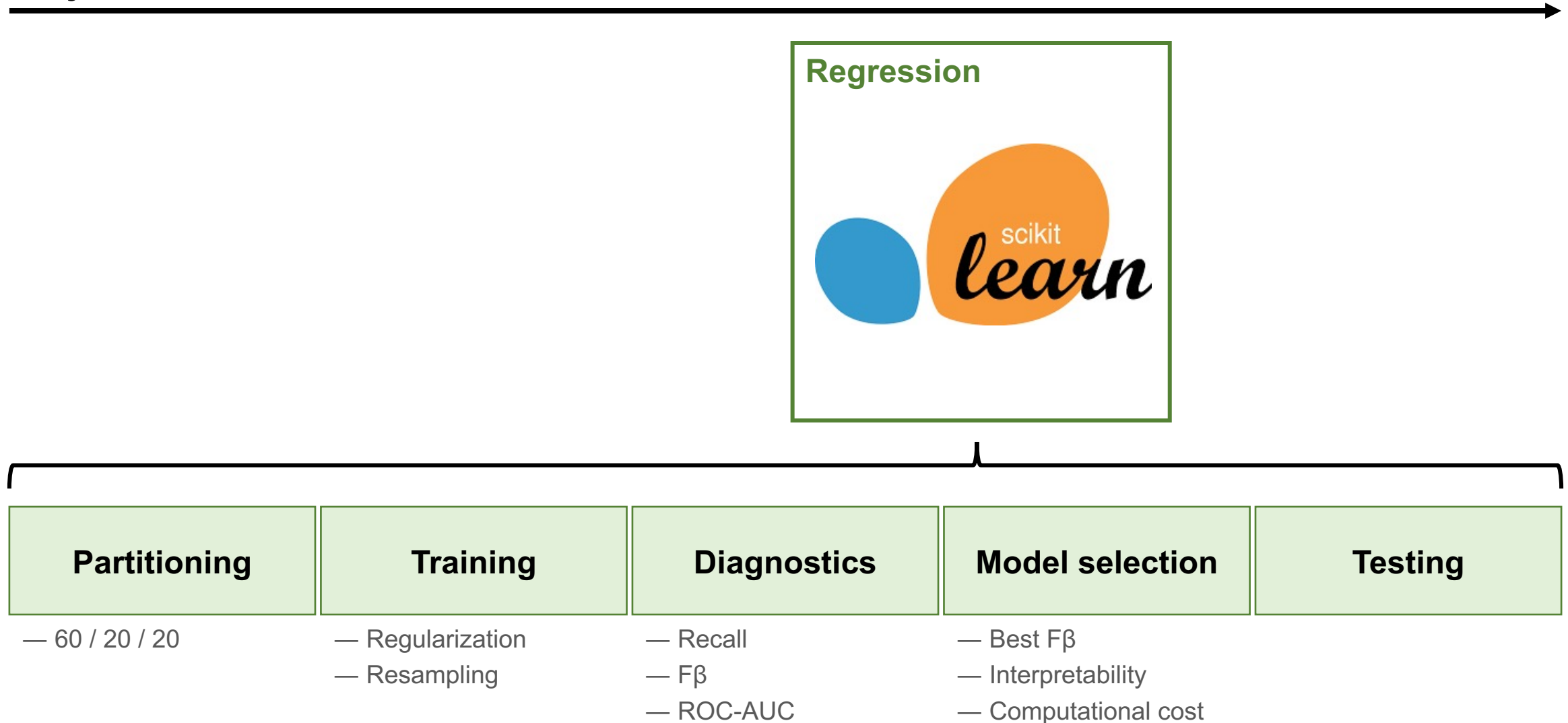
## Project workflow



- **252k observations**
  - Target class: loan default
  - Target distr.: ~10% default rate on sample
- **11 features**
  - Target: default/no default
  - Numerical: income, age, work experience, employment, time at residence
  - Categorical: marriage status, house ownership, car ownership, profession

# Methodology

## Project workflow



# Results

## Validation

### ▪ KNN:

- Recall: 0.81
- $F\beta$ : 0.72
- ROC: 0.88

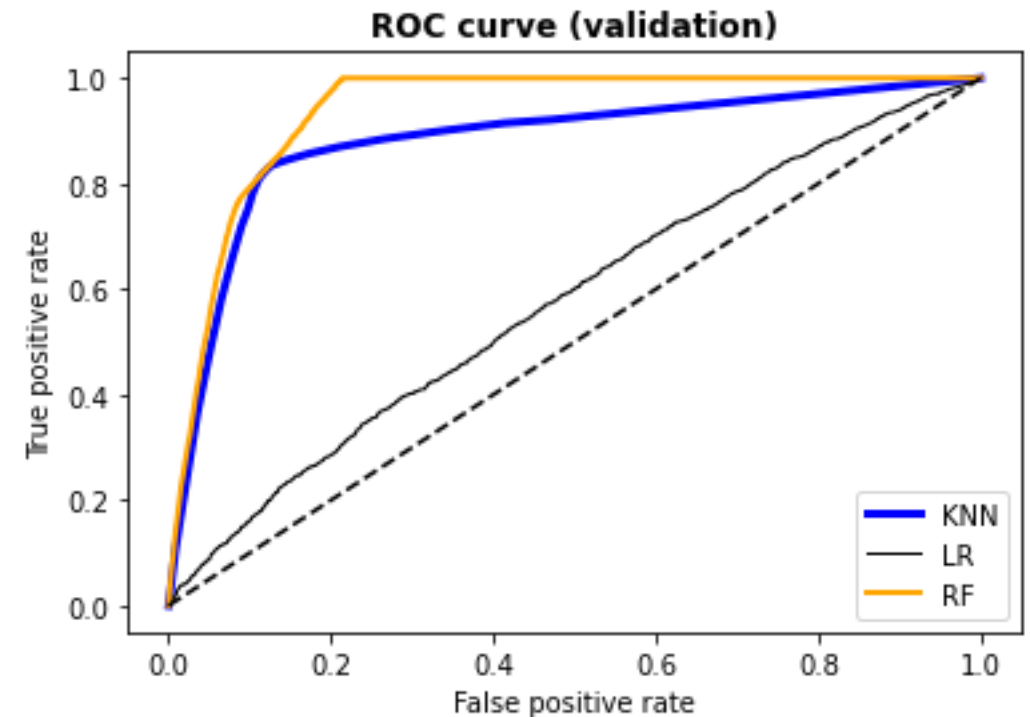
### ▪ Logistic Regression:

- Recall: 0.03
- $F\beta$ : 0.04
- ROC: 0.58

### ▪ Random Forest:

- Recall: 0.76
- $F\beta$ : 0.71
- ROC: 0.94

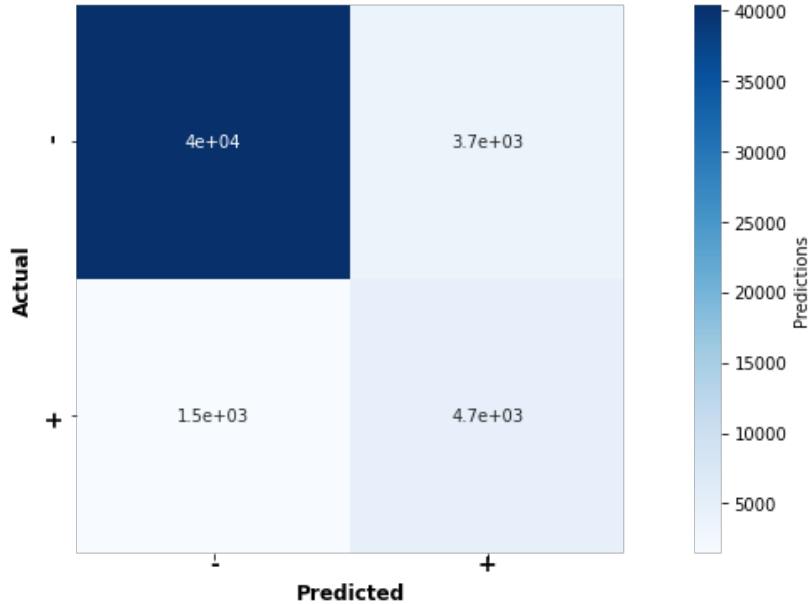
**Marginally worse recall/f-beta than KNN but acceptable, and will scale better with Ks/Ms more observations**



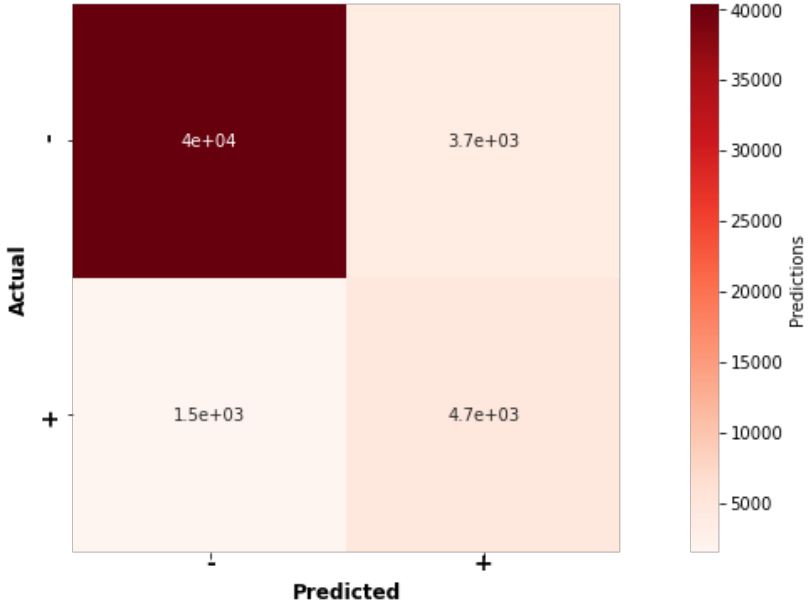
	model	recall	precision	f1	f-beta	roc
0	KNN	0.812893	0.501441	0.620266	0.723071	0.880015
1	Logistic Regression	0.032393	0.276860	0.058000	0.039341	0.577027
2	Random Forest	0.760516	0.557867	0.643617	0.709006	0.938439

# Conclusions

RF confusion matrix (validation classification)



RF confusion matrix (test classification)



RF results on test

Accuracy: 0.8956150793650793  
Recall: 0.7580361426515273  
Precision: 0.558435438265787  
F1: 0.6431042670103793  
F-beta: 0.7074626865671643  
ROC: 0.9378033040255929

The model holds reasonably stable on test, and could be a viable tool for credit relief outreach

# Further due diligence/future work



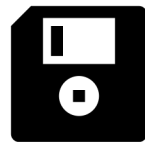
- **Parameter tuning**

- Decision thresholds
- KNN: neighbors, distance metric
- RF: depth, more estimators



- **More models to employ**

- Boosted trees
- Naïve Bayes



- **Incorporate more data**

- Monthly payment data
- More demographics

**Sam Reiff**

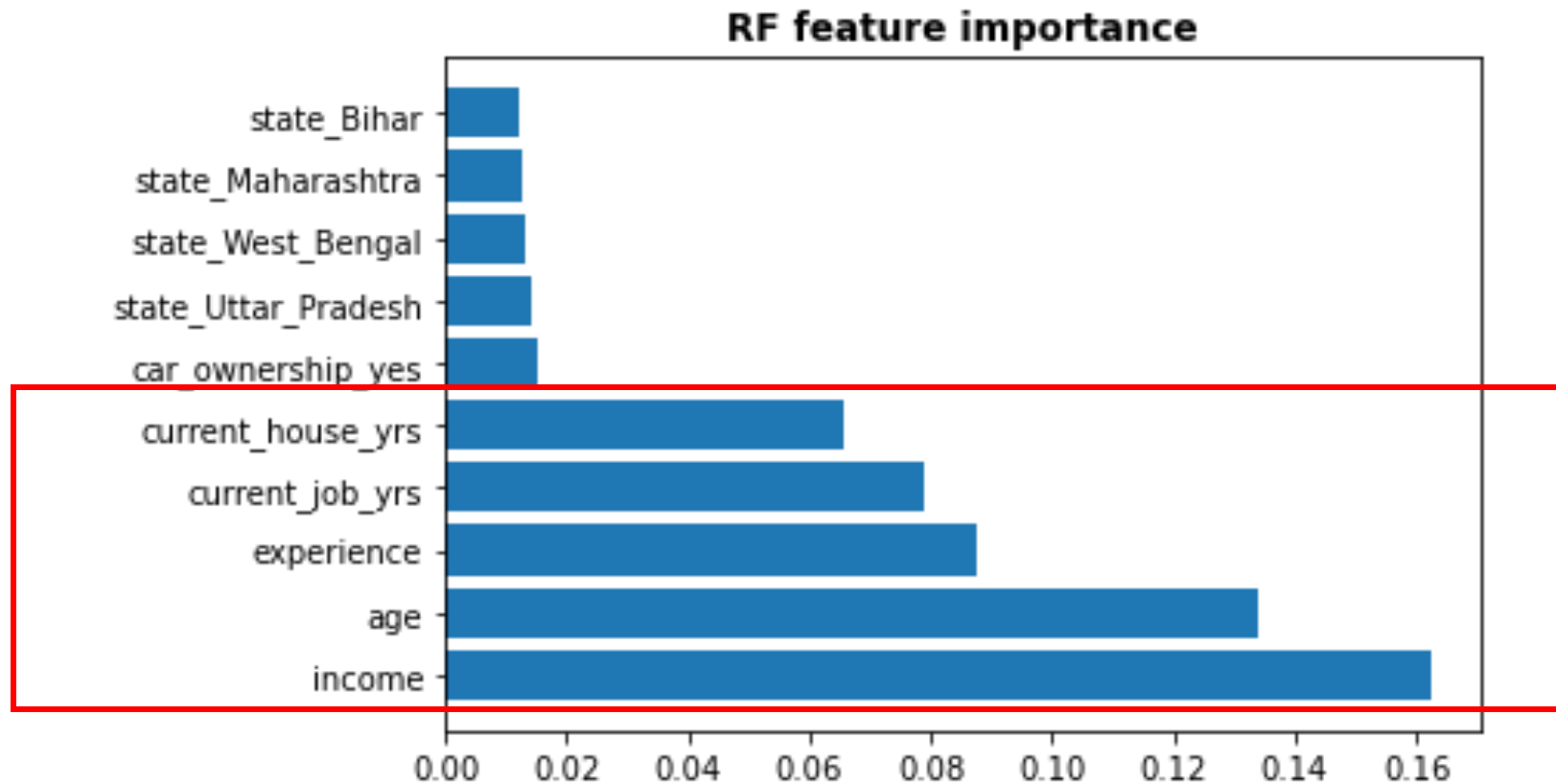
**reiff.sam@gmail.com**

**479.426.3700**



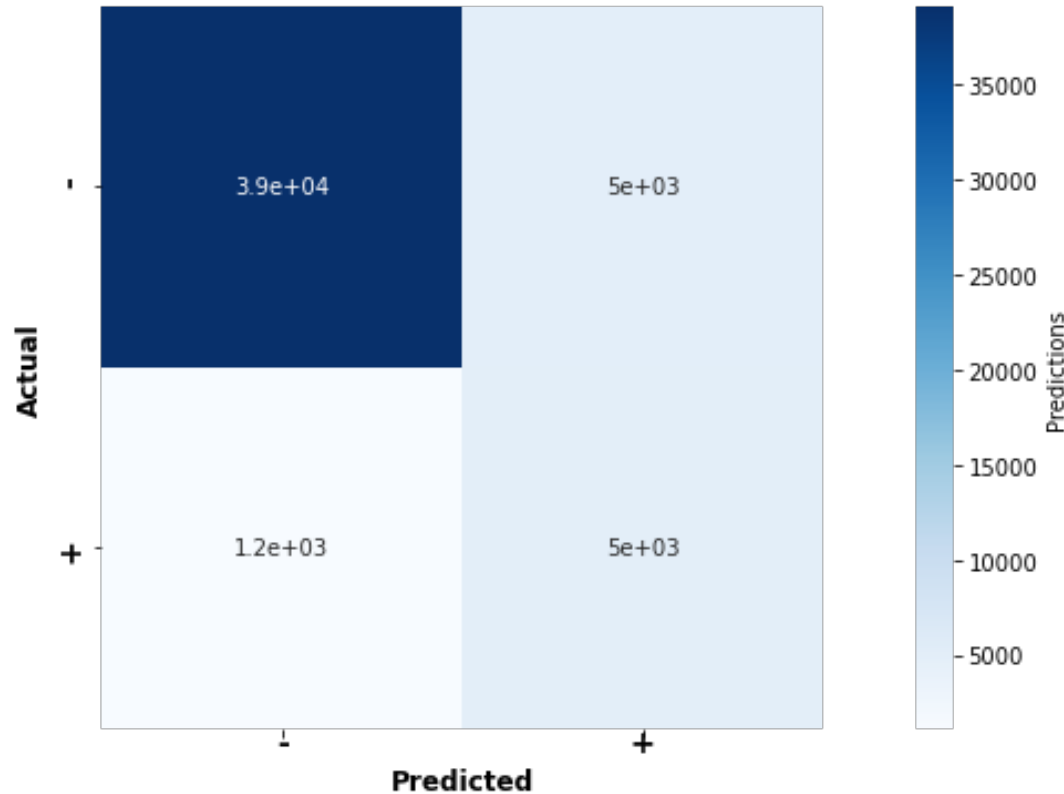
# Appendix

# Random Forest feature importance, top ten influential feats

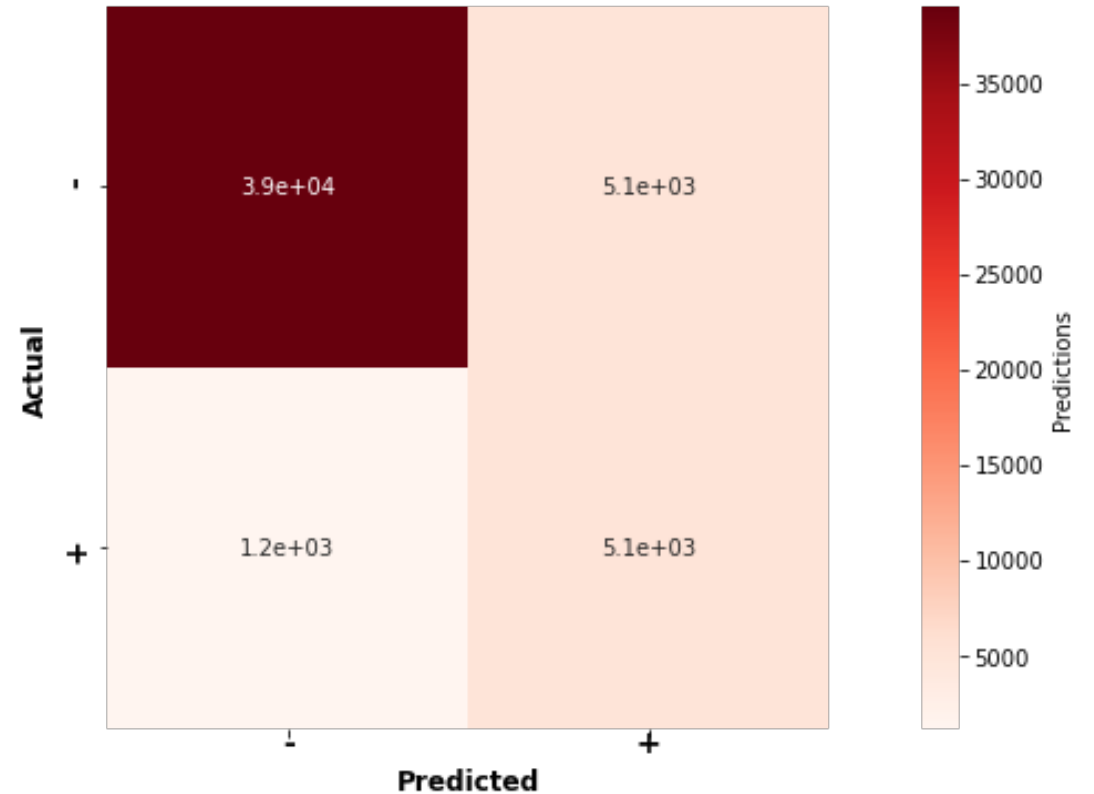


# KNN confusion matrix on validation and test

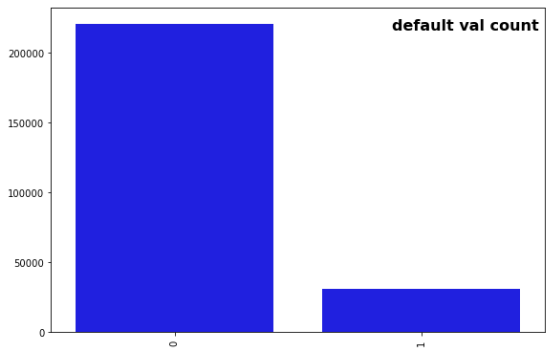
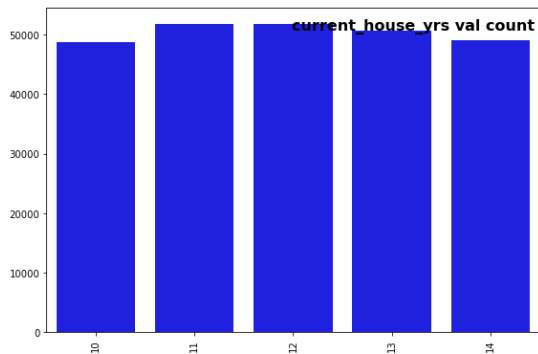
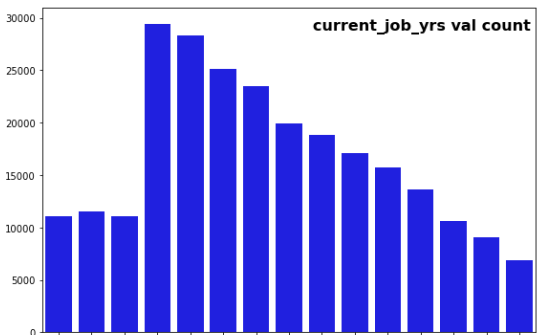
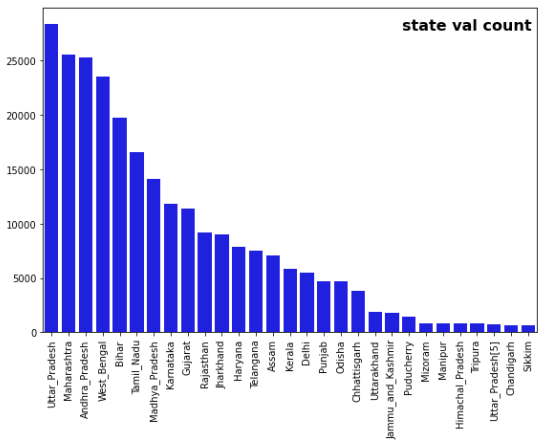
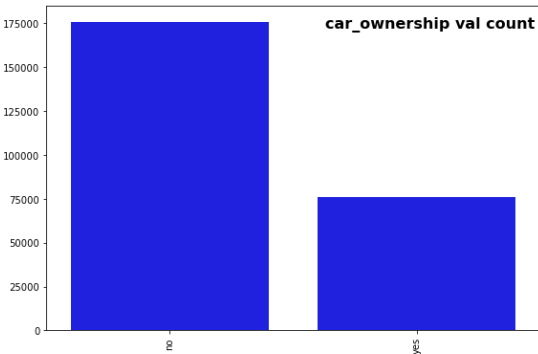
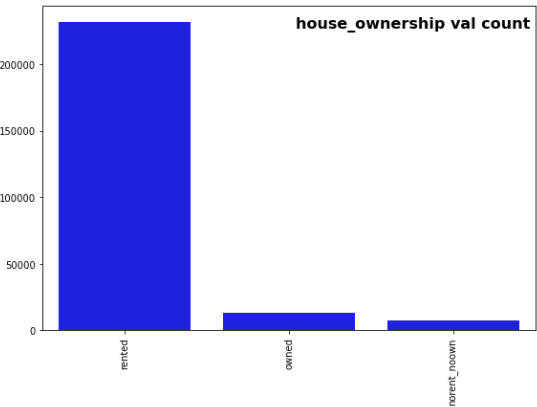
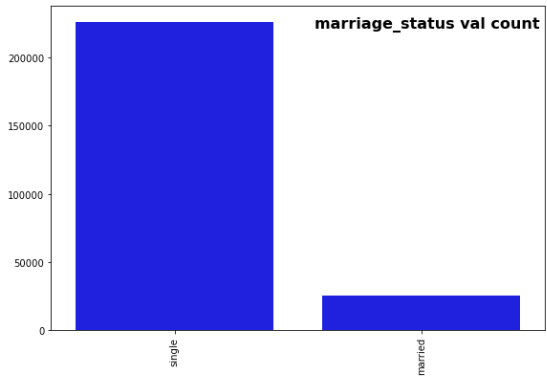
KNN confusion matrix (validation classification)



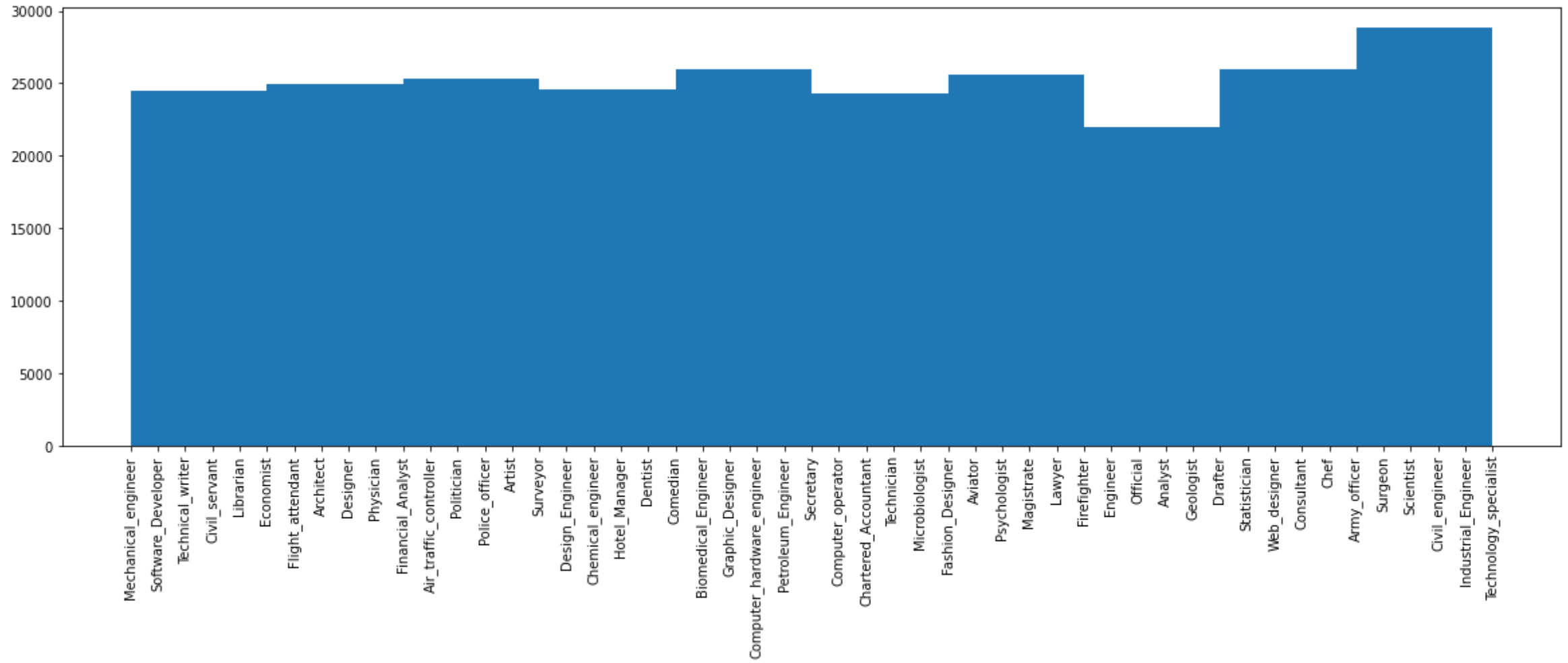
KNN confusion matrix (test classification)



# Feature value counts



# Profession feature distribution



# KNN validation results

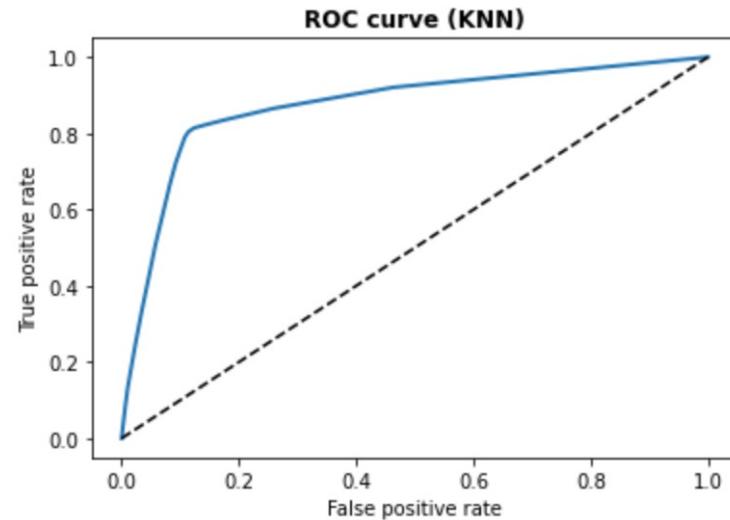
## KNN classification metrics with no resampling

---

KNN classification accuracy of: 0.8882539682539683  
KNN classification recall of: 0.41418211120064463  
KNN classification precision of: 0.5627326472520254  
KNN classification F1 of: 0.47716301522465654  
KNN classification F-beta (2) of: 0.4372681798073978

---

ROC AUC score = 0.8735908556312999



# KNN validation results

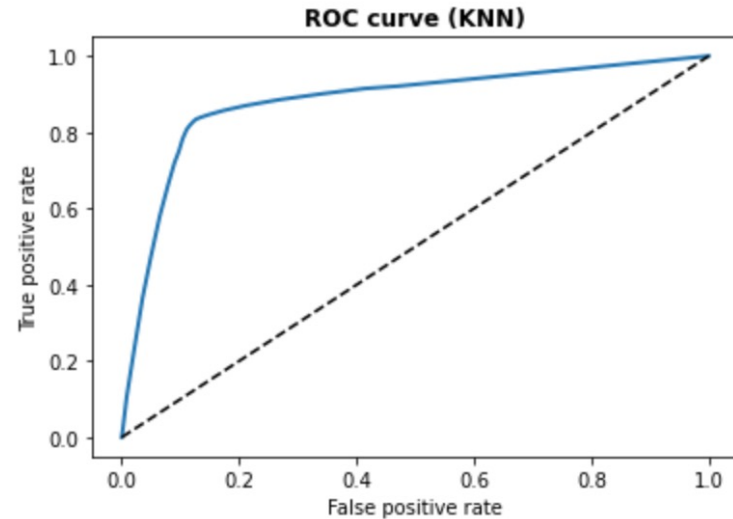
## KNN classification metrics with resampling

---

KNN classification accuracy of: 0.8774603174603175  
KNN classification recall of: 0.8128928283642224  
KNN classification precision of: 0.5014414951784472  
KNN classification F1 of: 0.6202656173143137  
KNN classification F-beta (2) of: 0.7230711889675736

---

ROC AUC score = 0.8800150931713429



# Logistic regression validation results

## Logistic Regression classification metrics with no resampling

---

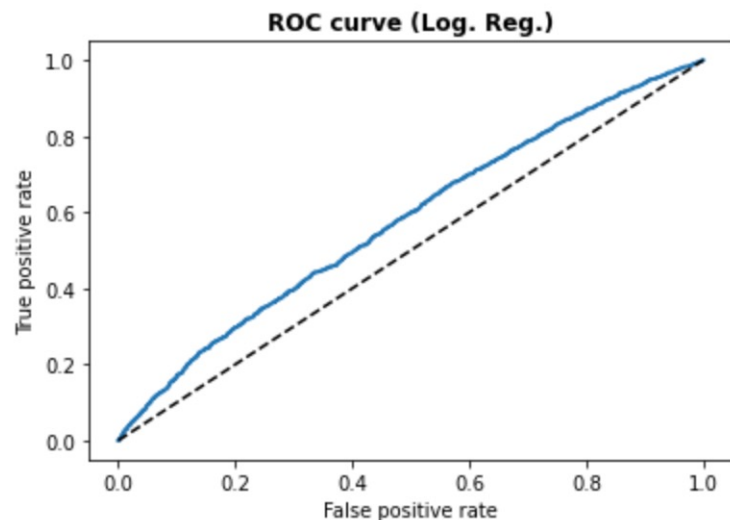
Logistic regression (hard preds) accuracy of: 0.8768849206349206  
Logistic regression (hard preds) recall of: 0.0  
Logistic regression (hard preds) precision of: 0.0  
Logistic regression (hard preds) F1 of: 0.0  
Logistic regression (hard preds) F-beta (2) of: 0.0

---

Logistic regression (soft preds with 0.17 decision threshold) accuracy of: 0.8340674603174603  
Logistic regression (soft preds with 0.17 decision threshold) recall of: 0.11845286059629331  
Logistic regression (soft preds with 0.17 decision threshold) precision of: 0.2025909592061742  
Logistic regression (soft preds with 0.17 decision threshold) F1 of: 0.149496593104851  
Logistic regression (soft preds with 0.17 decision threshold) F-beta (2) of: 0.12918307086614175  
Logistic regression (soft preds with 0.17 decision threshold) log loss of: 0.3689427109489983

---

ROC AUC score = 0.5768461945124708





# Logistic regression validation results

## Logistic Regression classification metrics with resampling

---

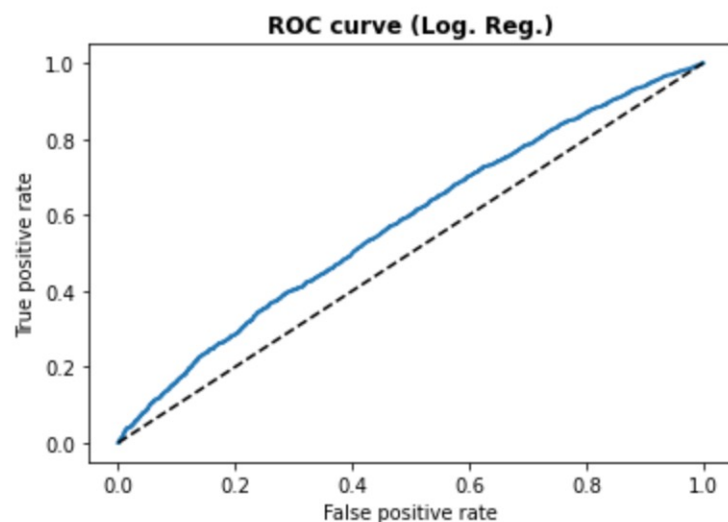
Logistic regression (hard preds) accuracy of: 0.8704563492063492  
Logistic regression (hard preds) recall of: 0.032393231265108784  
Logistic regression (hard preds) precision of: 0.2768595041322314  
Logistic regression (hard preds) F1 of: 0.05800028855864955  
Logistic regression (hard preds) F-beta (2) of: 0.0393407969936585

---

Logistic regression (soft preds with 0.17 decision threshold) accuracy of: 0.1252579365079365  
Logistic regression (soft preds with 0.17 decision threshold) recall of: 0.9988718775181306  
Logistic regression (soft preds with 0.17 decision threshold) precision of: 0.12327459326146625  
Logistic regression (soft preds with 0.17 decision threshold) F1 of: 0.21946426358373317  
Logistic regression (soft preds with 0.17 decision threshold) F-beta (2) of: 0.4126607899011958  
Logistic regression (soft preds with 0.17 decision threshold) log loss of: 0.5096175315499527

---

ROC AUC score = 0.5770271739987578



# Random forest validation results

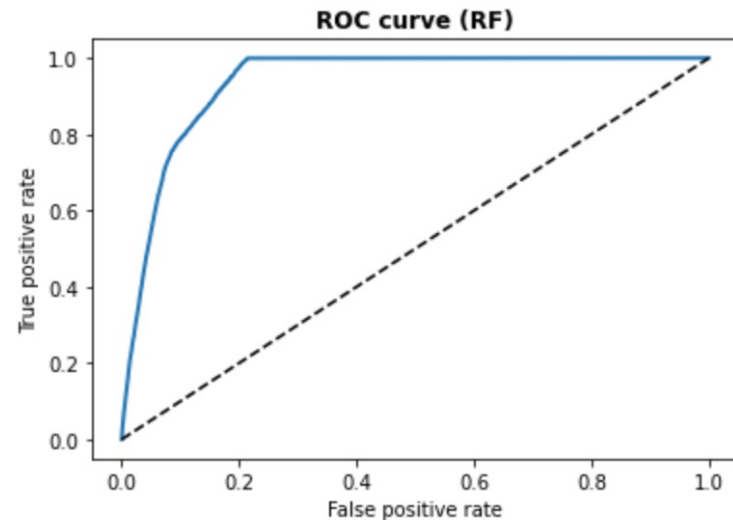
## Random Forest classification metrics with no resampling

---

Random forest classification accuracy of: 0.8998015873015873  
Random forest classification recall of: 0.5381144238517325  
Random forest classification precision of: 0.6045627376425855  
Random forest classification F1 of: 0.569406548431105  
Random forest classification (hard preds) F-beta (2) of: 0.5502092739676366

---

ROC AUC score = 0.9380972685425798



# Random forest validation results

## Random Forest classification metrics with resampling

---

Random forest classification accuracy of: 0.8963095238095238  
Random forest classification recall of: 0.7605157131345689  
Random forest classification precision of: 0.5578673602080624  
Random forest classification F1 of: 0.6436170212765957  
Random forest classification F-beta (2) of: 0.7090056792571892

---

ROC AUC score = 0.9384385951973339

