Credit Card Fraud Detection

Springboard Capstone Project

Chris J Mears - September 2020

The Problem

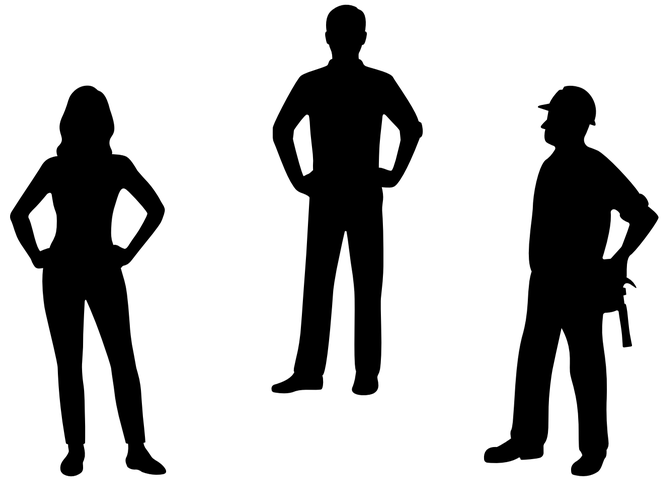
"Fraud losses worldwide reached $27.85 billion in 2018 and are projected to rise to $35.67 billion in five years and $40.63 billion in 10 years."

The Nilson Report

Who does this affect?

Banks

Lost Revenue & Work Hours



Businesses and Consumers

Frustration & Lost Time

How can we reduce fraud?

The Data

The Data

- Dataset from Kaggle
- Credit card transactions made in September 2013 by European cardholders
- Highly imbalanced
  - 492 fraud cases out of 284,807 transactions (0.173%)

The Data

- 28 PCA transformed features
  - Confidentiality
- 3 non-transformed features
  - Time (elapsed)
  - Amount
  - Class (fraud, non-fraud)

The Data

	Time	V1	V2	V3	V4	V5	...	V25	V26	V27	V28	Amount	Class
178698	123737.0	1.876980	-0.696291	0.450719	1.508030	-1.020571	...	-0.153727	-0.475249	0.113934	-0.023137	23.45	0
76266	56484.0	-2.905057	-2.553843	1.842890	-0.212129	0.480680	...	0.162048	0.701272	-0.250248	0.127260	427.64	0
51127	44807.0	1.095063	-0.302038	0.561085	-0.038436	-0.673402	...	0.008530	0.782091	-0.078727	0.015462	81.75	0
201147	133746.0	-2.739547	1.594012	-1.905468	0.576733	0.886780	...	-0.294596	-1.103543	-0.922379	-0.229078	79.00	0
73617	55206.0	0.990980	-0.365387	-0.414428	0.237689	-0.404407	...	0.473201	1.054140	-0.148832	0.019787	178.80	0
273023	165382.0	1.978600	0.680265	-1.001308	3.520812	1.053766	...	-0.110191	-0.339700	-0.036152	-0.040014	8.64	0
82712	59517.0	1.198267	0.265913	0.399505	0.631122	-0.465584	...	0.138864	0.066304	-0.030512	0.021475	1.98	0
112805	72813.0	-0.510041	0.801484	1.955988	1.010750	0.430620	...	0.381213	-0.195755	-0.342545	-0.403932	57.00	0
30261	35844.0	-3.222500	-3.641709	2.577789	-0.929482	2.402566	...	0.685703	-0.106645	0.173078	0.184205	12.48	0
79333	57972.0	-2.558141	-2.292401	1.450960	0.108837	-2.089814	...	0.022859	0.123339	0.944169	-0.198398	445.00	0
123561	76978.0	1.050040	-1.777259	1.310847	-0.410554	-1.618252	...	0.513019	0.173519	0.089214	0.022859	123.00	0
14975	26291.0	-0.451848	0.927968	0.816701	1.963048	-0.920190	...	-1.007315	-0.324020	-0.216883	0.132860	44.03	0
255807	157410.0	-0.213966	0.990962	-1.084647	-0.310215	0.605261	...	-0.458031	0.511308	-0.202088	-0.015042	10.55	0
46909	42985.0	-4.075975	0.963031	-5.076070	4.955963	-0.161437	...	-0.304987	-0.106089	1.899714	0.511462	1.00	1
235167	148278.0	2.046638	-0.099353	-1.203877	0.207330	0.119362	...	-0.270674	0.202065	-0.073970	-0.073668	1.98	0

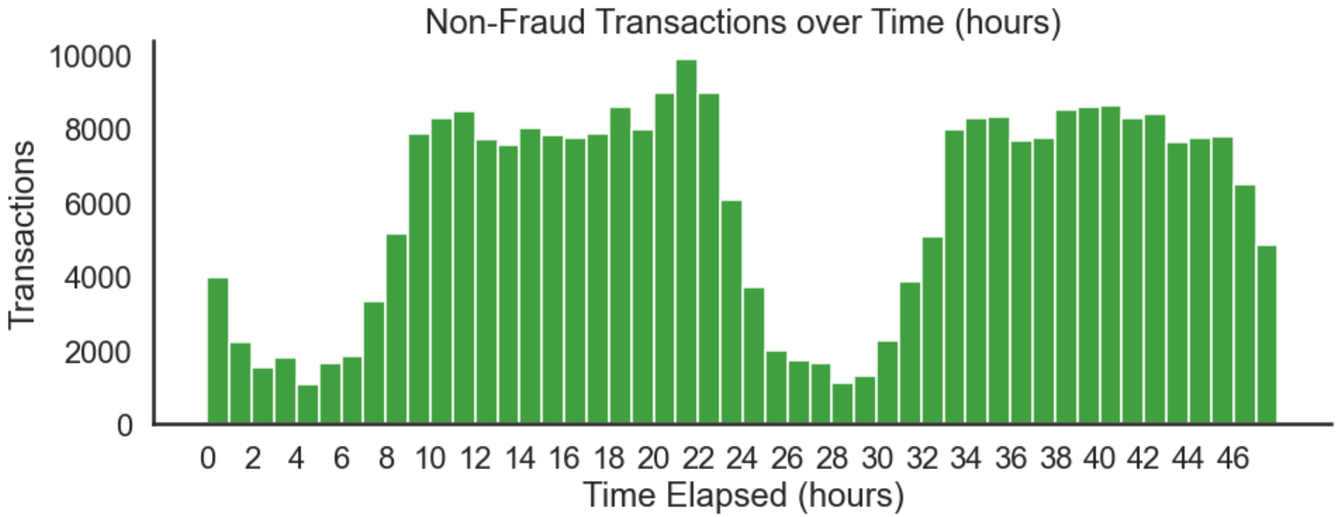
The Data

	Time	V1	V2	V3	V4	V5	...	V25	V26	V27	V28	Amount	Class
178698	123737.0	1.876980	-0.696291	0.450719	1.508030	-1.020571	...	-0.153727	-0.475249	0.113934	-0.023137	23.45	0
76266	56484.0	-2.905057	-2.553843	1.842890	-0.212129	0.480680	...	0.162048	0.701272	-0.250248	0.127260	427.64	0
51127	44807.0	1.095063	-0.302038	0.561085	-0.038436	-0.673402	...	0.008530	0.782091	-0.078727	0.015462	81.75	0
201147	133746.0	-2.739547	1.594012	-1.905468	0.576733	0.886780	...	-0.294596	-1.103543	-0.922379	-0.229078	79.00	0
73617	55206.0	0.990980	-0.365387	-0.414428	0.237689	-0.404407	...	0.473201	1.054140	-0.148832	0.019787	178.80	0
273023	165382.0	1.978600	0.680265	-1.001308	3.520812	1.053766	...	-0.110191	-0.339700	-0.036152	-0.040014	8.64	0
82712	59517.0	1.198267	0.265913	0.399505	0.631122	-0.465584	...	0.138864	0.066304	-0.030512	0.021475	1.98	0
112805	72813.0	-0.510041	0.801484	1.955988	1.010750	0.430620	...	0.381213	-0.195755	-0.342545	-0.403932	57.00	0
30261	35844.0	-3.222500	-3.641709	2.577789	-0.929482	2.402566	...	0.685703	-0.106645	0.173078	0.184205	12.48	0
79333	57972.0	-2.558141	-2.292401	1.450960	0.108837	-2.089814	...	0.022859	0.123339	0.944169	-0.198398	445.00	0
123561	76978.0	1.050040	-1.777259	1.310847	-0.410554	-1.618252	...	0.513019	0.173519	0.089214	0.022859	123.00	0
14975	26291.0	-0.451848	0.927968	0.816701	1.963048	-0.920190	...	-1.007315	-0.324020	-0.216883	0.132860	44.03	
255807	157410.0	-0.213966	0.990962	-1.084647	-0.310215	0.605261	...	-0.458031	0.511308	-0.202088	-0.015042	10.55	
46909	42985.0	-4.075975	0.963031	-5.076070	4.955963	-0.161437	...	-0.304987	-0.106089	1.899714	0.511462	1.00	1
235167	148278.0	2.046638	-0.099353	-1.203877	0.207330	0.119362	...	-0.270674	0.202065	-0.073970	-0.073668	1.98	0

Fraud!

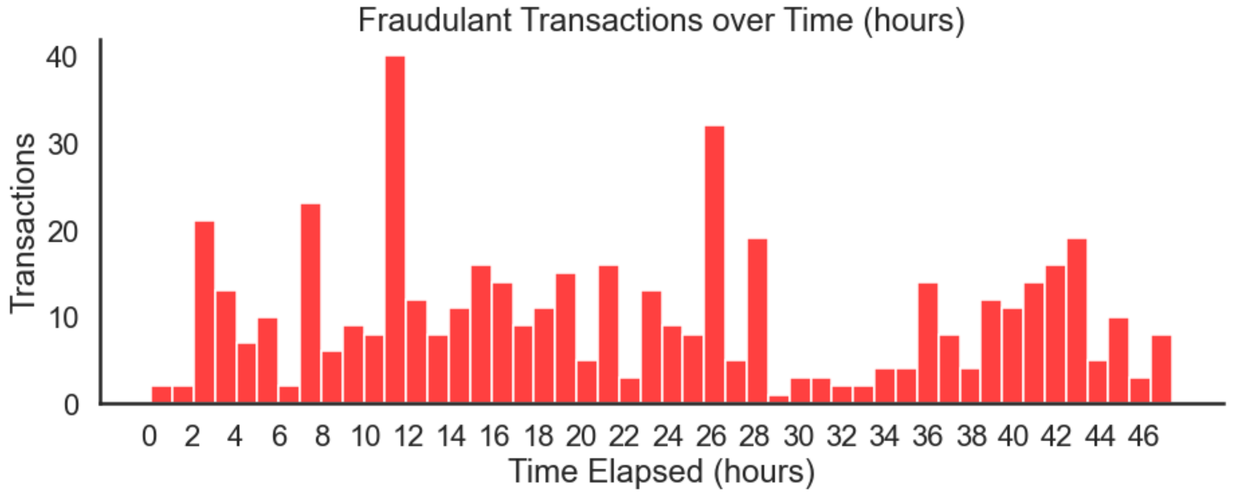
Exploratory Data Analysis

EDA - Transactions over Time



- Cyclic
- Time starts at midnight?

- Slightly cyclic; more random
- Far less frequent transactions



EDA - Amount

	Mean	Standard Deviation	Max
Non-Fraud	88.29	250.10	25691.16
Fraud	122.21	256.68	2125.87

EDA - Class

0.173%

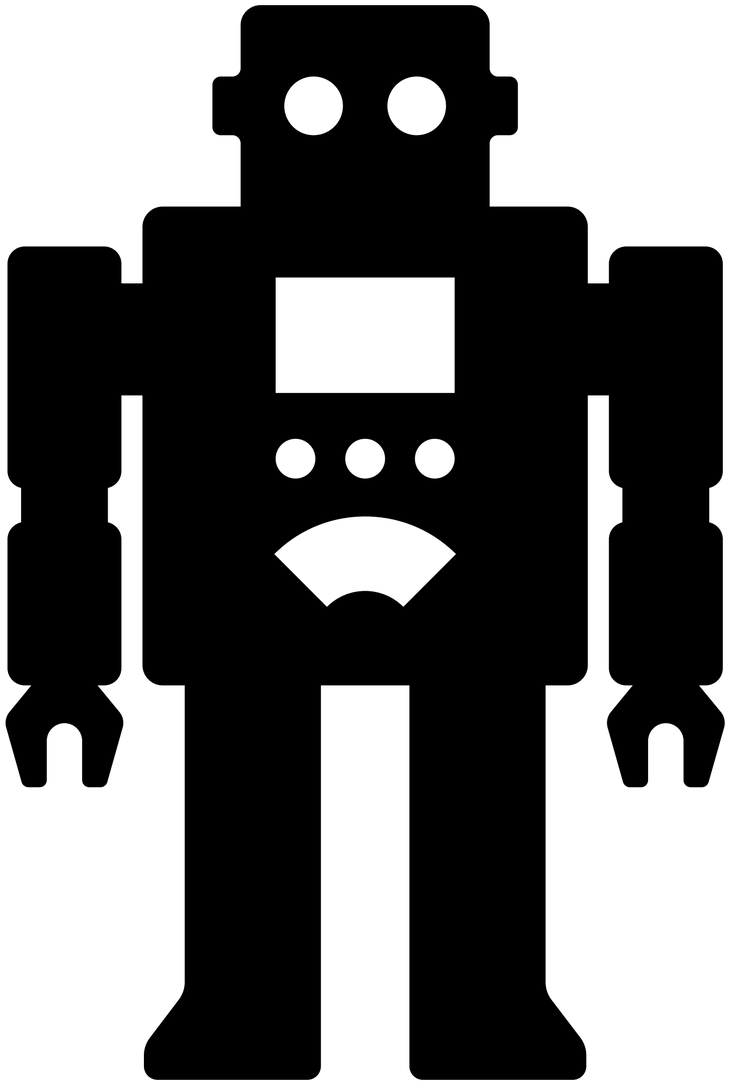
492 fraud cases out of 284,807 transactions

Model Selection

Model Selection

Supervised Learning

- Classification model
  - Binary (0 for non-fraud; 1 for fraud)
- Highly imbalanced
- Tools: scikit-learn, imblearn



Model Selection

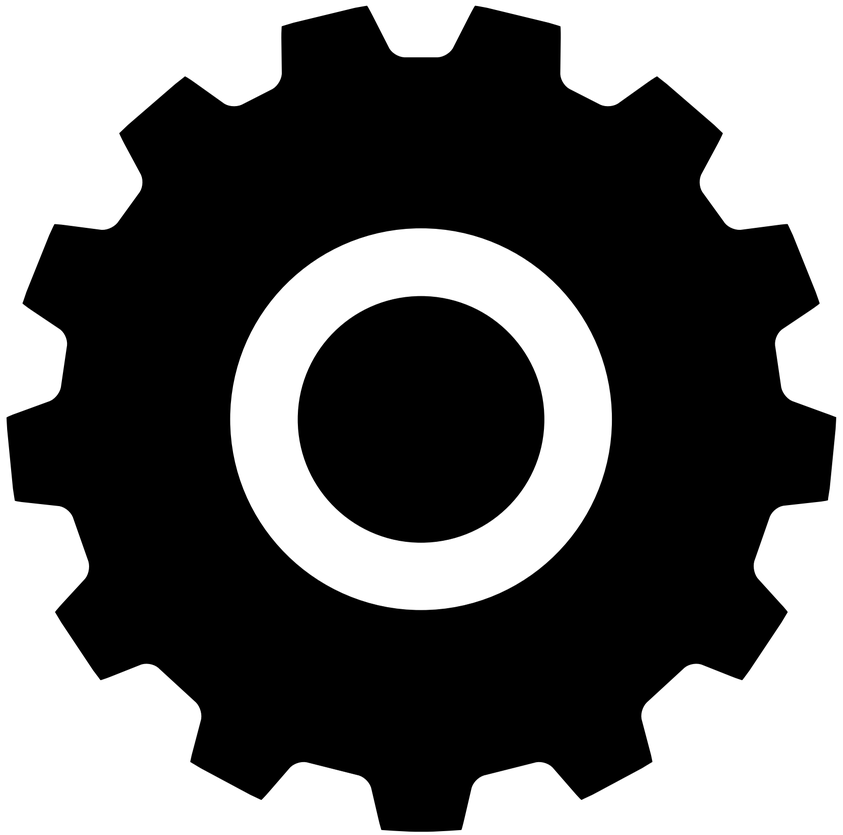
Three Models

1. Logistical Regression
2. Random Forest Classifier
3. Support Vector Machines with the Radial Basis Function (RBF) kernel

Feature Engineering
&
Evaluation

Feature Engineering

- Scaling
  - StandardScaler for SVM only
- Not much other else needed
  - Clean dataset
  - PCA transformation did heavy lifting



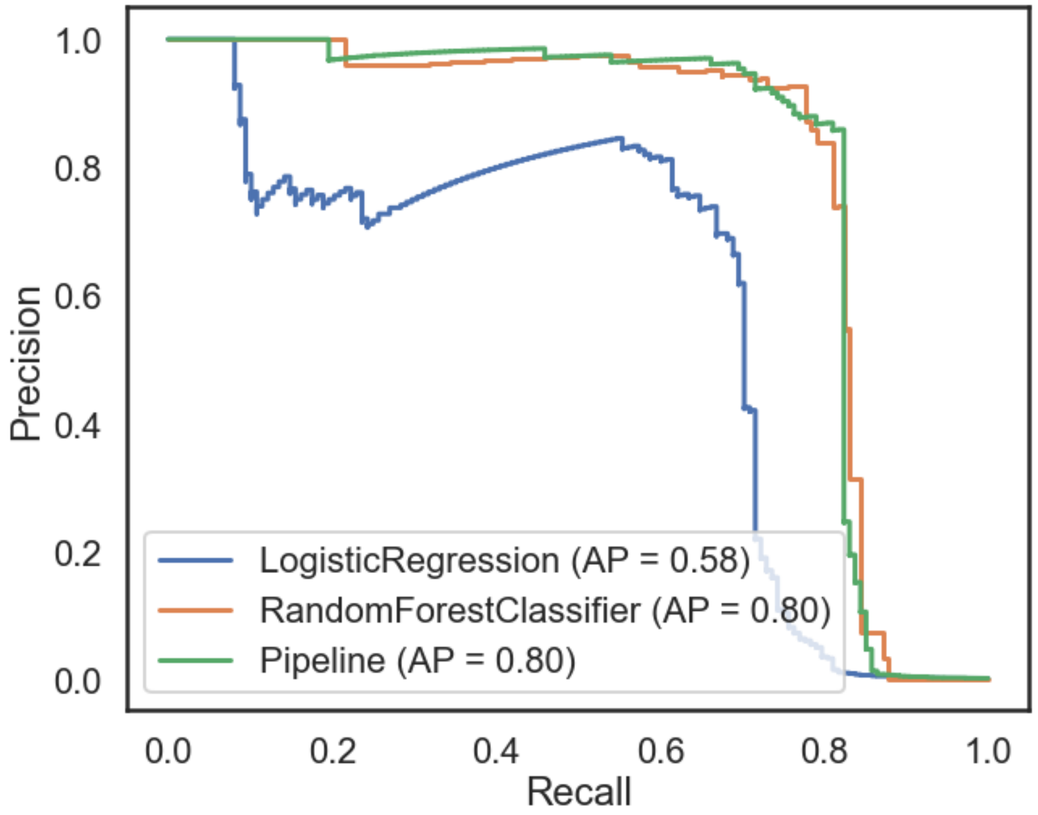
Pre-Processing

- 70%/30% train/test split with stratification

	Fraud Class %	Fraud Count	Non-Fraud Count
Full Dataset	0.173%	492	284807
Training Set	0.173%	344	199364
Testing Set	0.173%	148	85443

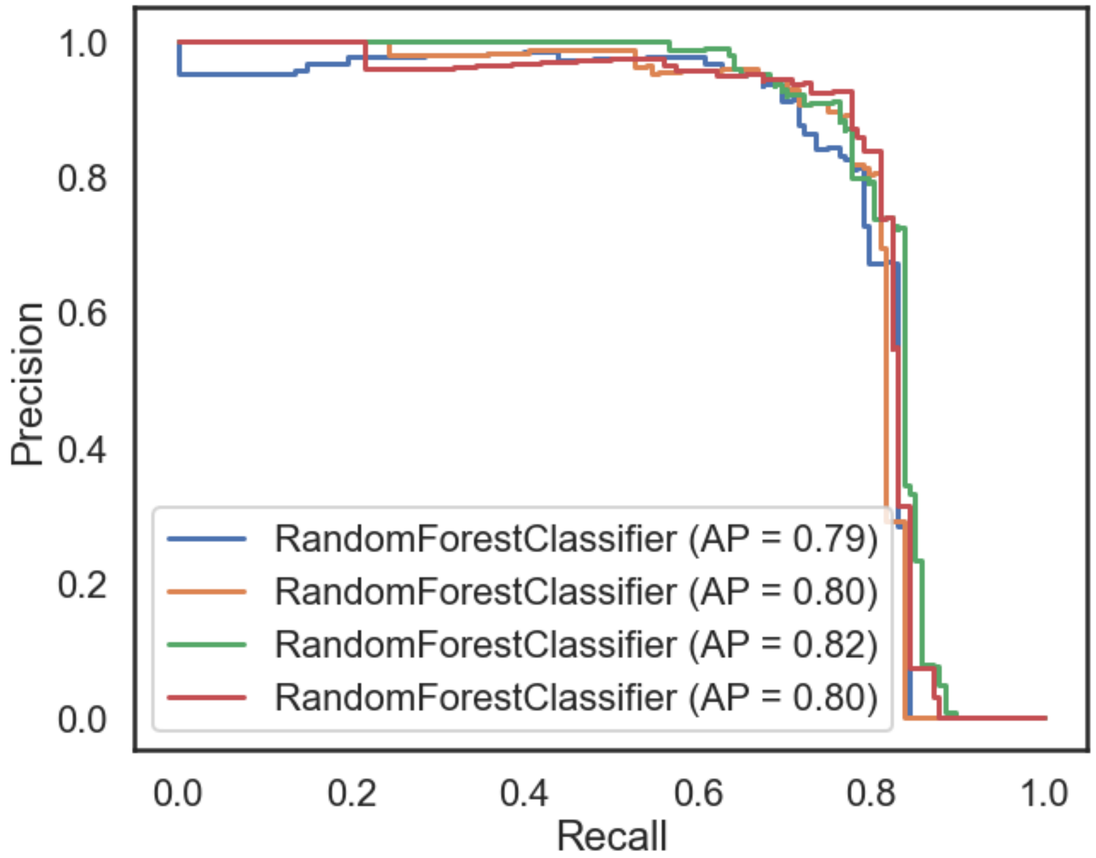
Evaluation Results

Model	P/R AUC	F1	MCC
Logistic Regression	0.5803	0.6906	0.6916
Random Forest Classifier	**0.819**	**0.8365**	**0.8429**
SVM w/ RBF Kernel	0.804	0.7469	0.7668



Upsampling Experiments

Model	Fraud Count	Non-Fraud Count	P/R AUC
Upsampling Minority Class 1:1	199020	199020	**0.8024**
Upsampling Minority Class 1:2	99510	199020	**0.8221**
SMOTE Upsampling 1:1	199020	199020	**0.8216**
SMOTE Upsampling 1:2	99510	199020	**0.8094**



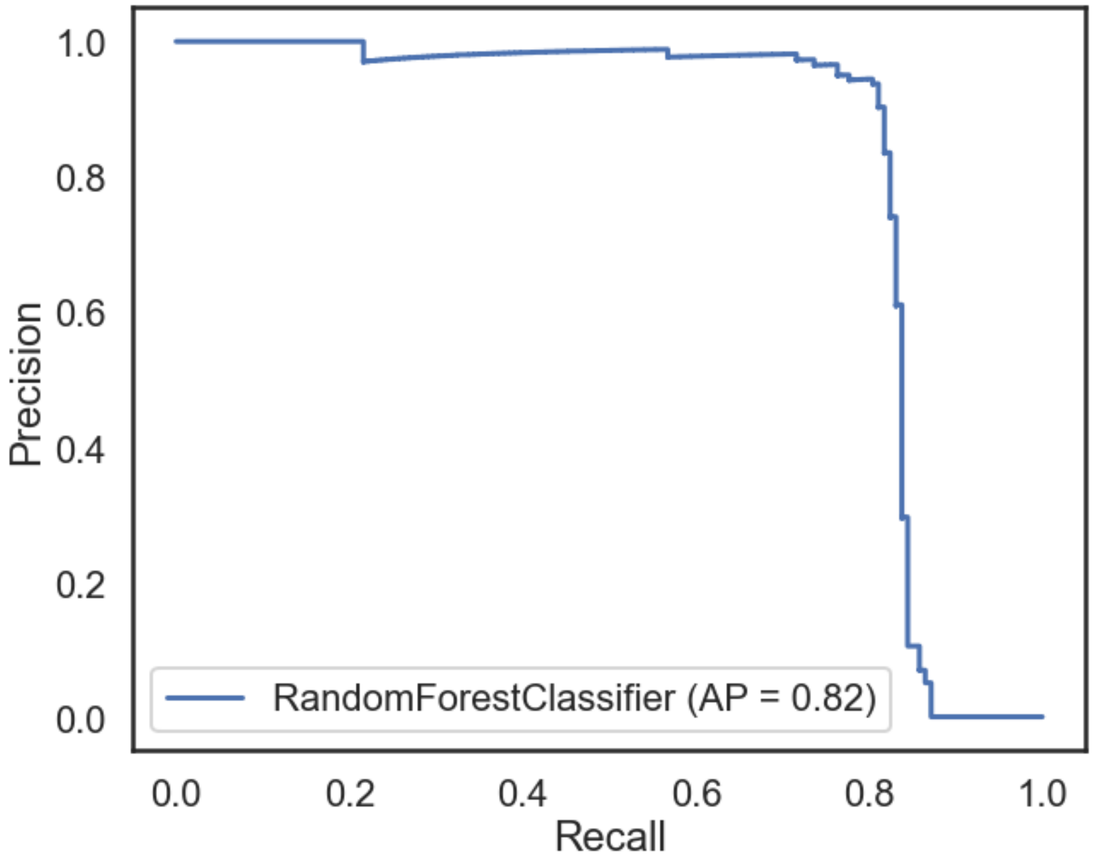
Final Model

RandomForestClassifier

Final Model

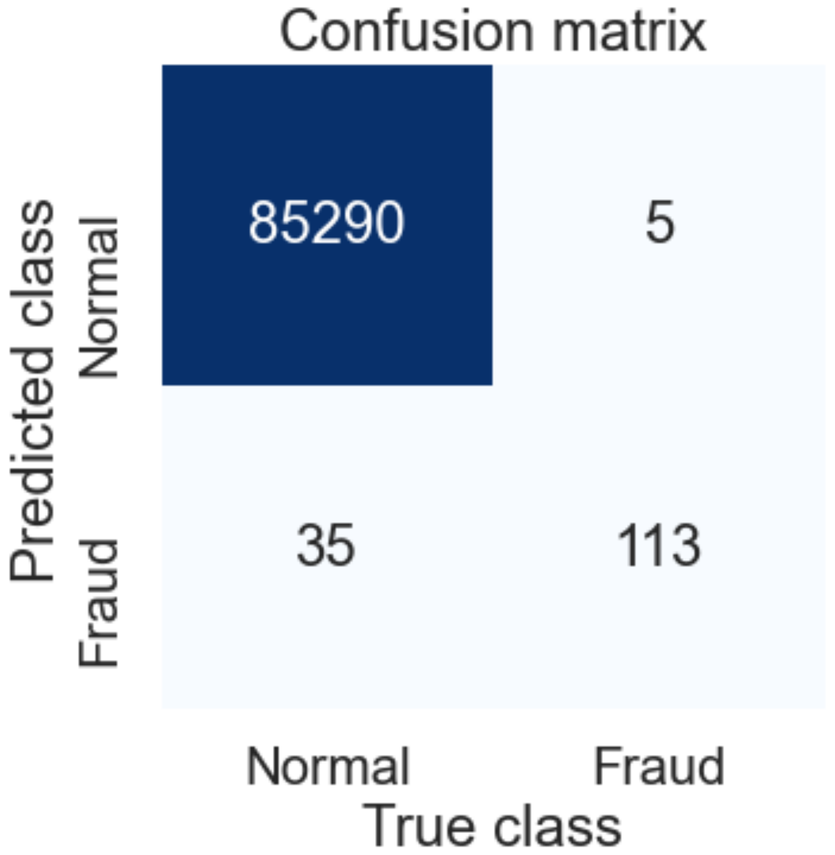
RandomForestClassifier

- **PR / AUC = 0.8253**
- **Hyperparameters**
  - Bootstrap = False
  - Max Depth = 80
  - Max Features = Auto
  - Min Samples Leaf = 1
  - Min Samples Split = 10
  - Number of Estimators = 94



Conclusions

- RandomForestClassifier is decent
- Upsampling didn't work well in this case
- The data is a little old
- A good start for more experimentation



Further Research

- More models (Gaussian Naive Bayes, Gradient Boosting, Extremely Randomized Trees)
- Cross Validation with training / testing sets
- Utilize imblearn library
- Real-time detection with neural network?

Questions?