BANK OF AMERICA

Credit Card Fraudulent Detection

Kartik Garg Neo Liu

Zhiyi Zhao Michelle Tan

Agenda

- **Initial stage**
  - Business Case Value Evaluation
  - Check The Source Data
- **Data Exploration**
  - Explortary Data Analysis
  - Data Cleaning
  - Data Modeling
- **Algorithm**
  - Logistic Regression
  - Random Forest
  - Light Gradient Boosting Machine
  - Deep Learning
- **Summary**
  - Model Selection
  - Main Summary
  - Updated Value Evaluation

BANK OF AMERICA

Value Estimation

Current State

- Revenue - **$22.8B**
- Cost - **$14.4B**
- Current fraud detection algorithm - **60%**

Future State

- Increase fraud detection precision to 75%

Gaps

- **Increase revenue**
  - Reinvesting **$3.4B** to grow market share
  - **Value $3.4B**
- **Decrease costs**
  - Current Cost – **$14.4B**
  - At **75%** efficiency cost will be – **$13.2B**
  - **Value $2.2B**
- **Execute Both**
  - **$1.1B Save cost**
  - **$1.1B reinvest**

BANK OF AMERICA

Check The Source Data

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278

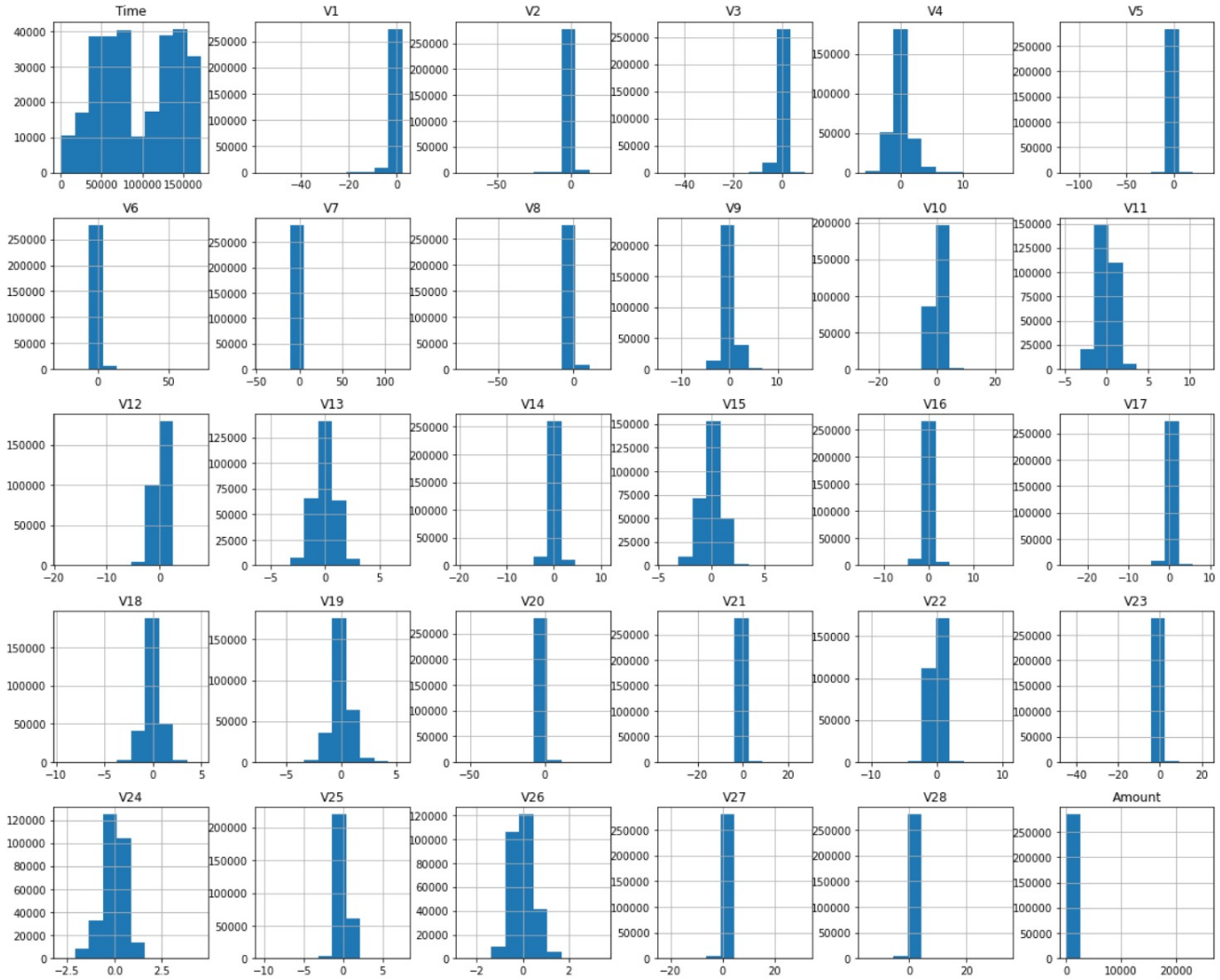
V23	V24	V25	V26	V27	V28	Amount	Class
-0.110474	0.066928	0.128539	-0.189115	0.133558	-0.021053	149.62	0
0.101288	-0.339846	0.167170	0.125895	-0.008983	0.014724	2.69	0
0.909412	-0.689281	-0.327642	-0.139097	-0.055353	-0.059752	378.66	0
-0.190321	-1.175575	0.647376	-0.221929	0.062723	0.061458	123.50	0
-0.137458	0.141267	-0.206010	0.502292	0.219422	0.215153	69.99	0

	Time	V1	V2	V3	V4	V5	V6	V7	V8
count	284807.000000	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05
mean	94813.859575	3.918649e-15	5.682686e-16	-8.761736e-15	2.811118e-15	-1.552103e-15	2.040130e-15	-1.698953e-15	-1.893285e-16
std	47488.145955	1.958696e+00	1.651309e+00	1.516255e+00	1.415869e+00	1.380247e+00	1.332271e+00	1.237094e+00	1.194353e+00
min	0.000000	-5.640751e+01	-7.271573e+01	-4.832559e+01	-5.683171e+00	-1.137433e+02	-2.616051e+01	-4.355724e+01	-7.321672e+01
25%	54201.500000	-9.203734e-01	-5.985499e-01	-8.903648e-01	-8.486401e-01	-6.915971e-01	-7.682956e-01	-5.540759e-01	-2.086297e-01
50%	84692.000000	1.810880e-02	6.548556e-02	1.798463e-01	-1.984653e-02	-5.433583e-02	-2.741871e-01	4.010308e-02	2.235804e-02
75%	139320.500000	1.315642e+00	8.037239e-01	1.027196e+00	7.433413e-01	6.119264e-01	3.985649e-01	5.704361e-01	3.273459e-01
max	172792.000000	2.454930e+00	2.205773e+01	9.382558e+00	1.687534e+01	3.480167e+01	7.330163e+01	1.205895e+02	2.000721e+01

V28	Amount	Class
2.848070e+05	284807.000000	284807.000000
-1.217809e-16	88.349619	0.001727
3.300833e-01	250.120109	0.041527
-1.543008e+01	0.000000	0.000000
-5.295979e-02	5.600000	0.000000
1.124383e-02	22.000000	0.000000
7.827995e-02	77.165000	0.000000
3.384781e+01	25691.160000	1.000000

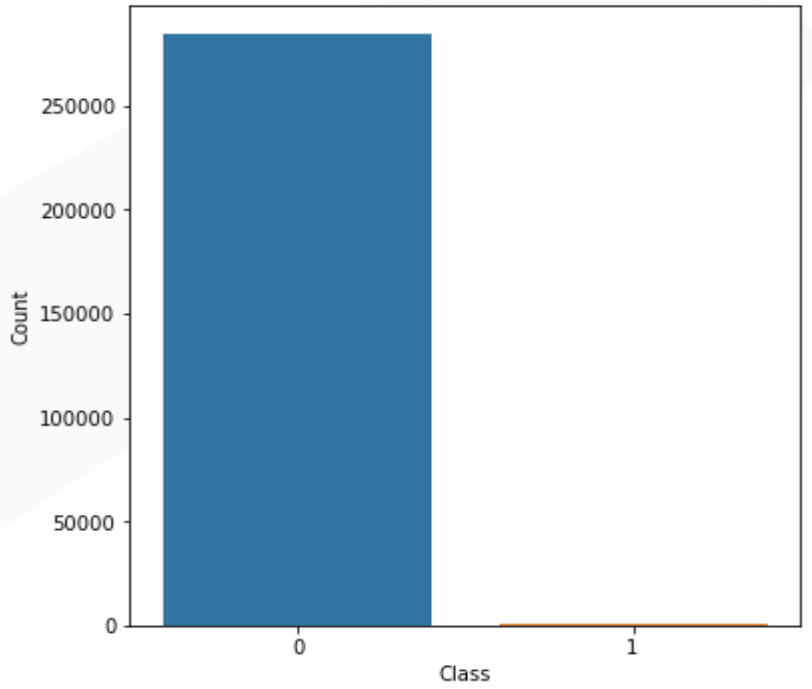
BANK OF AMERICA

Exploratory Data Analysis - Variable Distribution



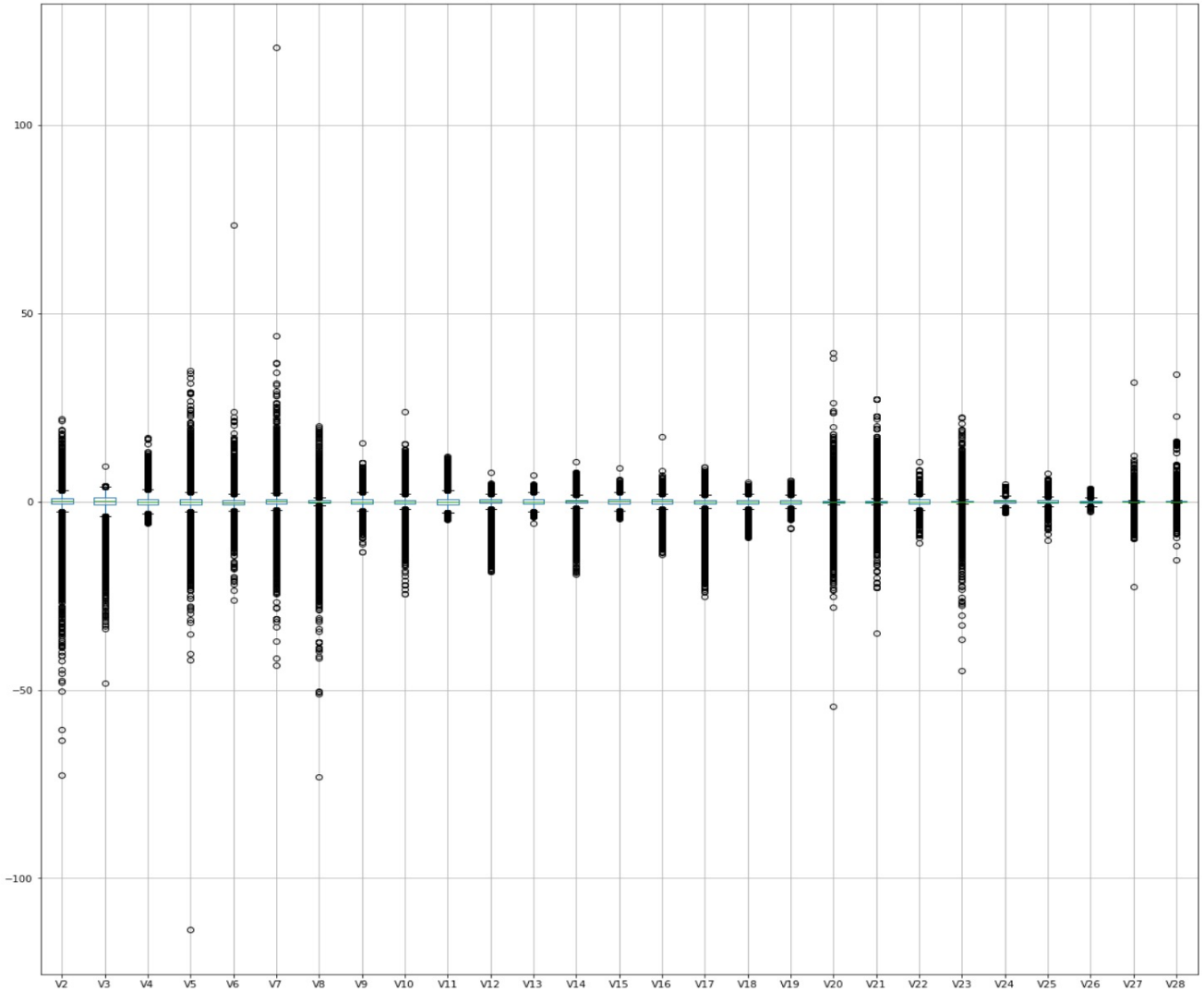
From the figure, we can observe that

1). Most variables are not normally distributed
2). The variables have different scales
3). Severe imbalanced dataset



BANK OF AMERICA

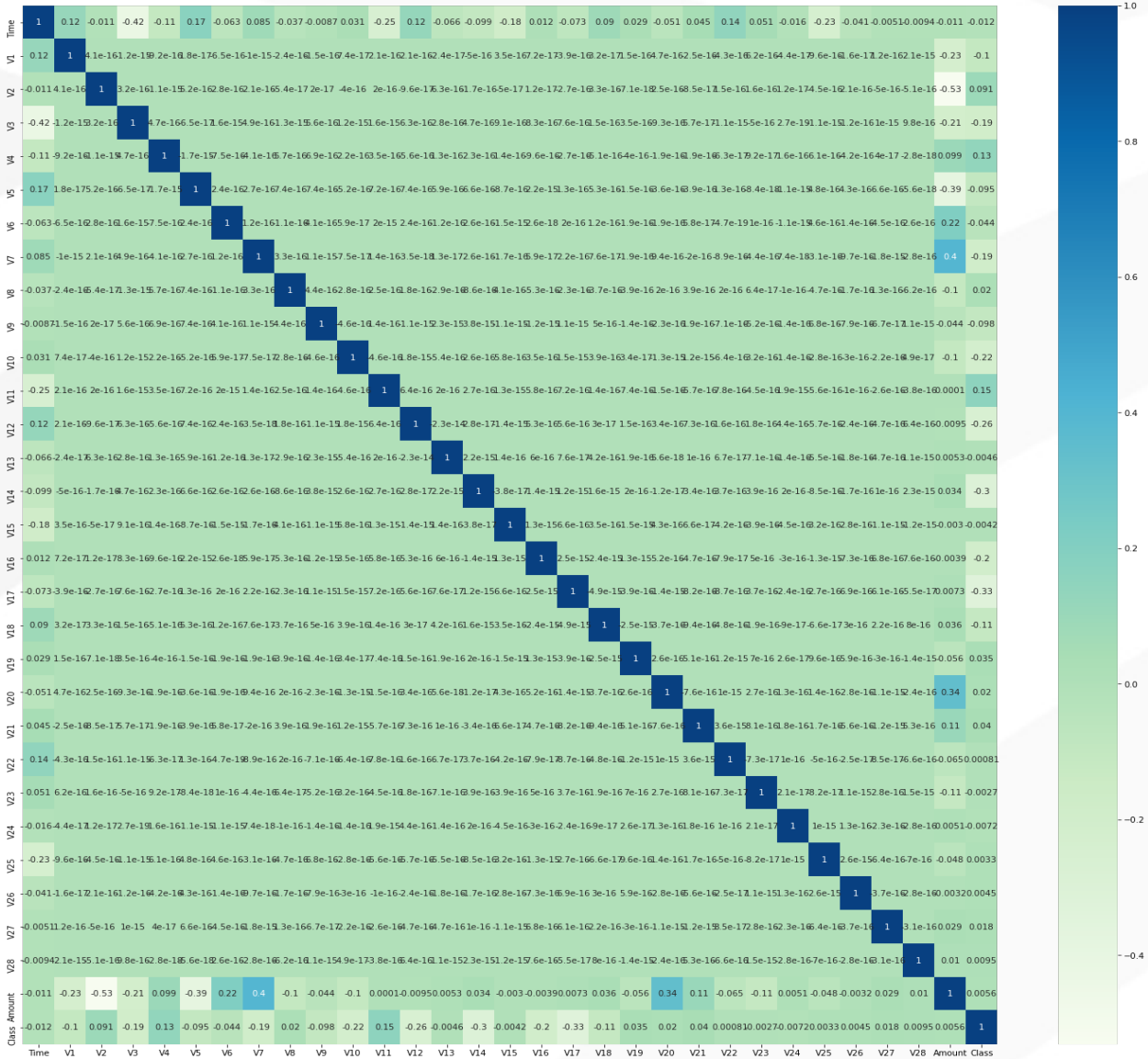
Exploratory Data Analysis - Outlier Visualization



We can spot outliers exist in different variables from the boxplot

BANK OF AMERICA

Exploratory Data Analysis - Correlation Between Variables



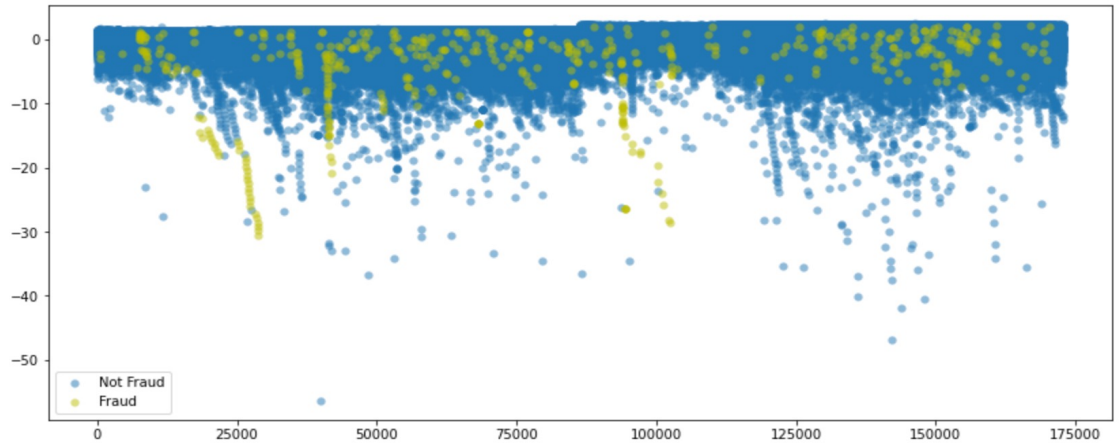
We can find some relatively outstanding correlation between variables

1). V2 and Amount are negatively correlated (-0.53)
2). V5 and Amount are negatively correlated (-0.39)
3). V7 and Amount are positively correlated (0.4)

BANK OF AMERICA

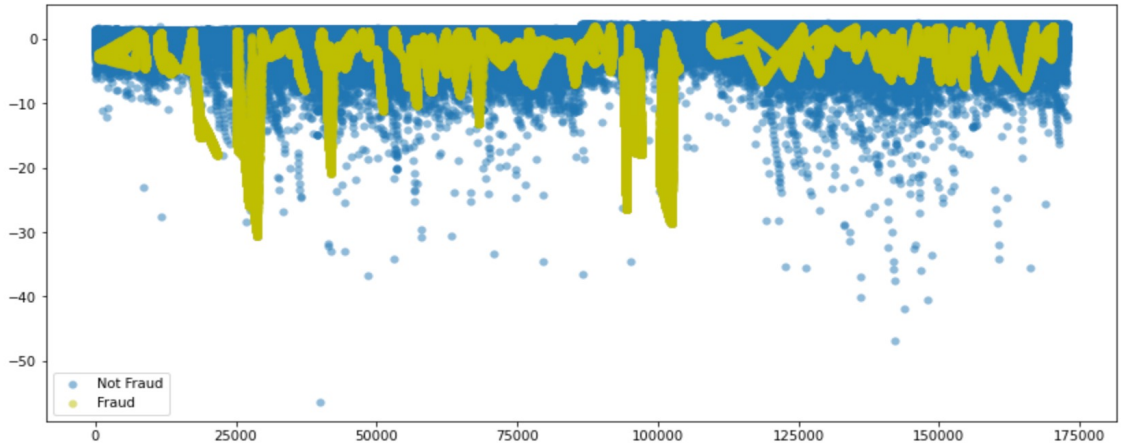
Exploratory Data Analysis - Oversampling Visualization

SMOTE



Before

- Imbalanced

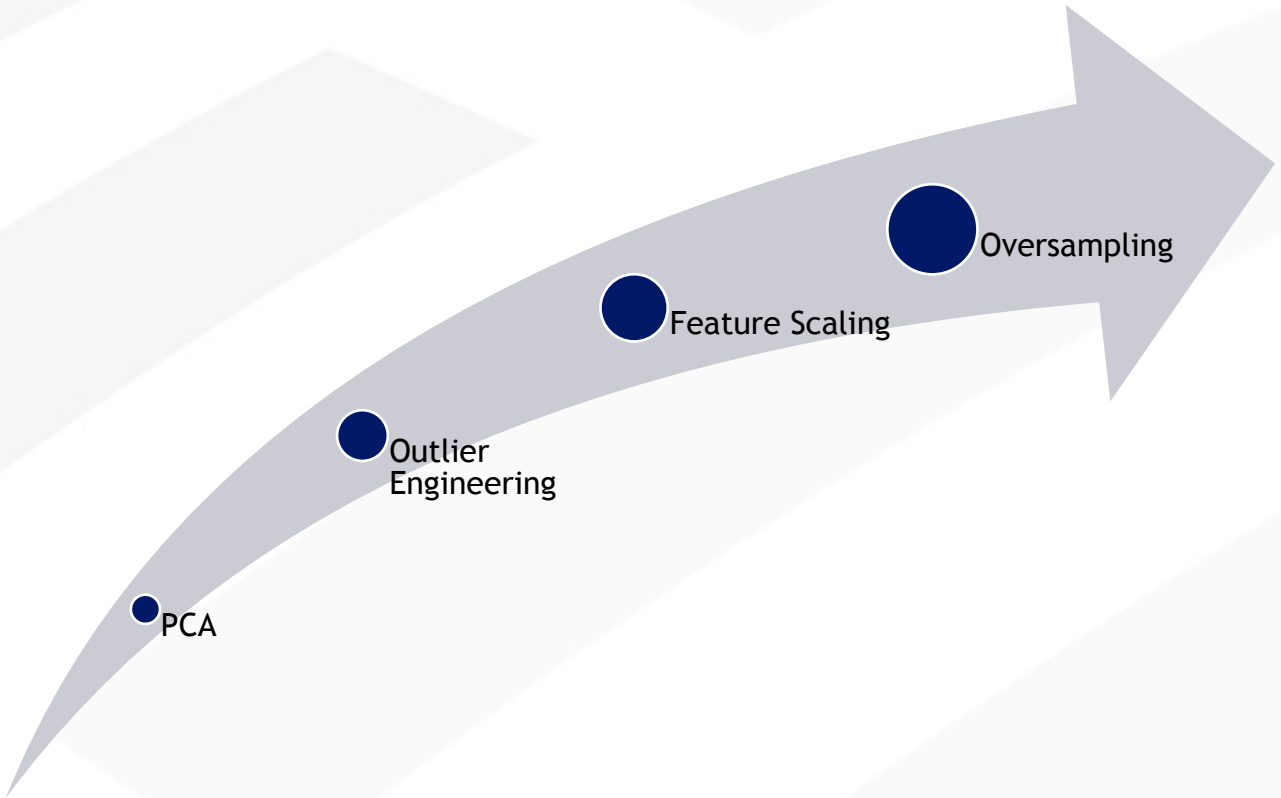


After

- Almost doubled the size of our data especially minority class by generating generic samples
- With better result after scaling

BANK OF AMERICA

Data Cleaning



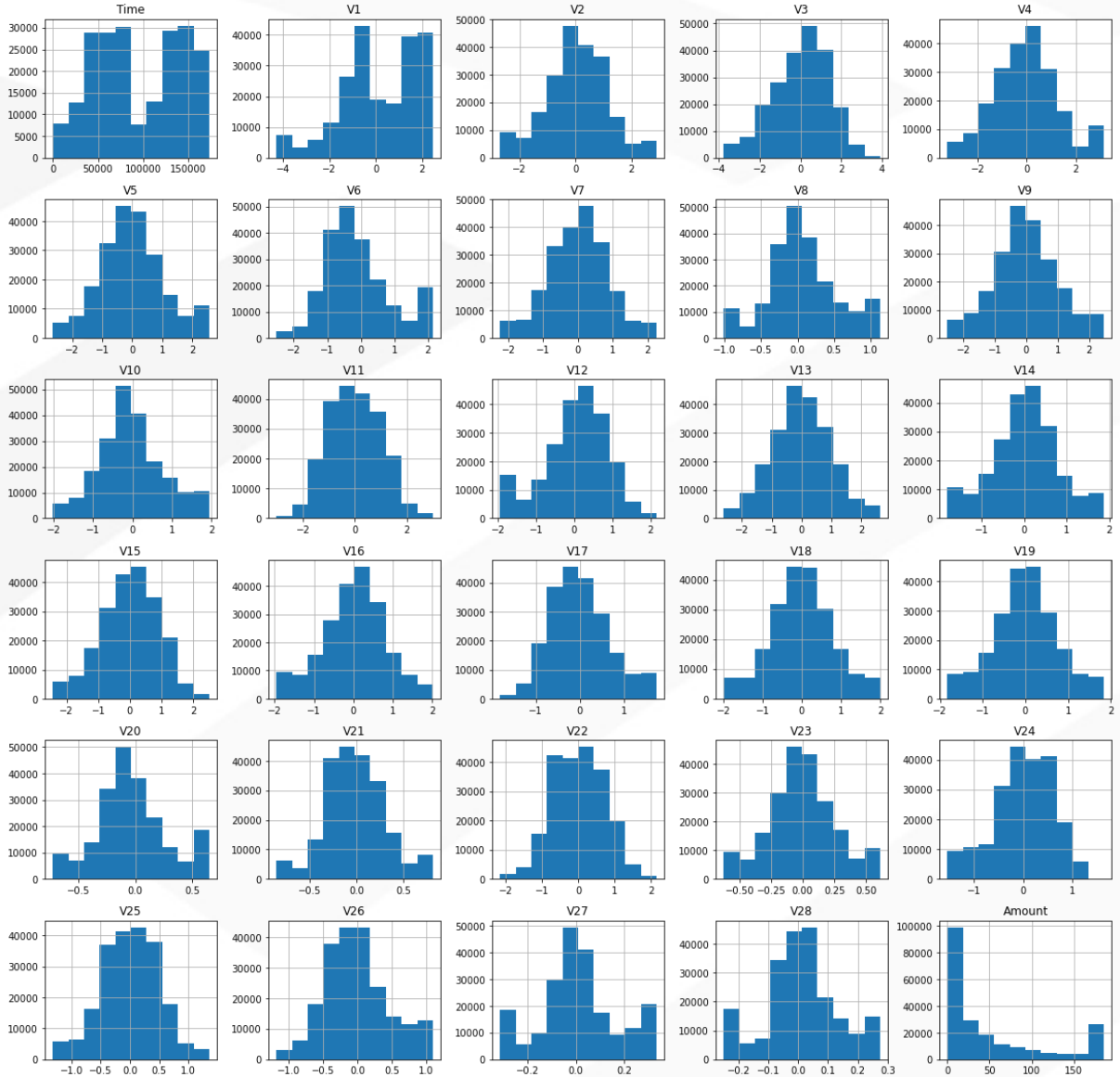
- Principal components obtained with PCA due to confidentiality

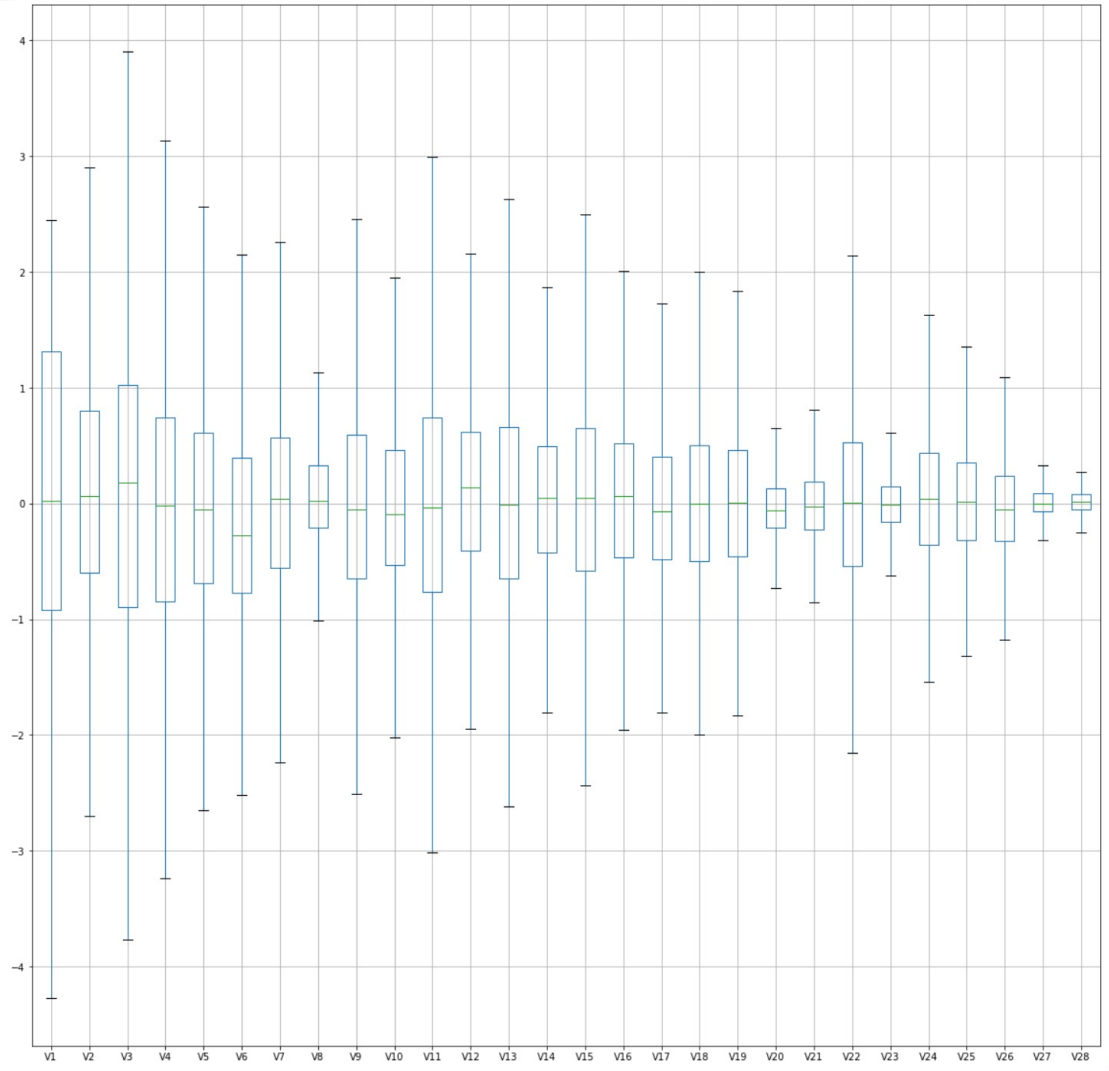
- Outlier Engineering – interquantile range
- Feature Scaling - Standardization

- Oversampling to achieve balanced dataset (SMOTE)

BANK OF AMERICA

Data Cleaning - Output





BANK OF AMERICA

Data Modeling - Conventional Rule based system

Not Machine Learning

- made up of a set of rule
- Predictive Stage

Pro

- easy to understand

Con

- manual work, time consuming and might not be able to identify edge cases

Success Metrics:

- Confusion Matrix
- Precision, Recall, F1 Score

Confusion Matrix

		Flagged Fraud	
		0	1
Actual Fraud	Non-Fraud	283089	1226
	Fraud	322	170

	Precision	Recall	F1 Score
Non-Fraud	1.00	1.00	1.00
Fraud	0.35	0.12	0.18

BANK OF AMERICA

Logistic Regression

Supervised Machine Learning

- Adding classification by modeling the probability of a certain event

Pro

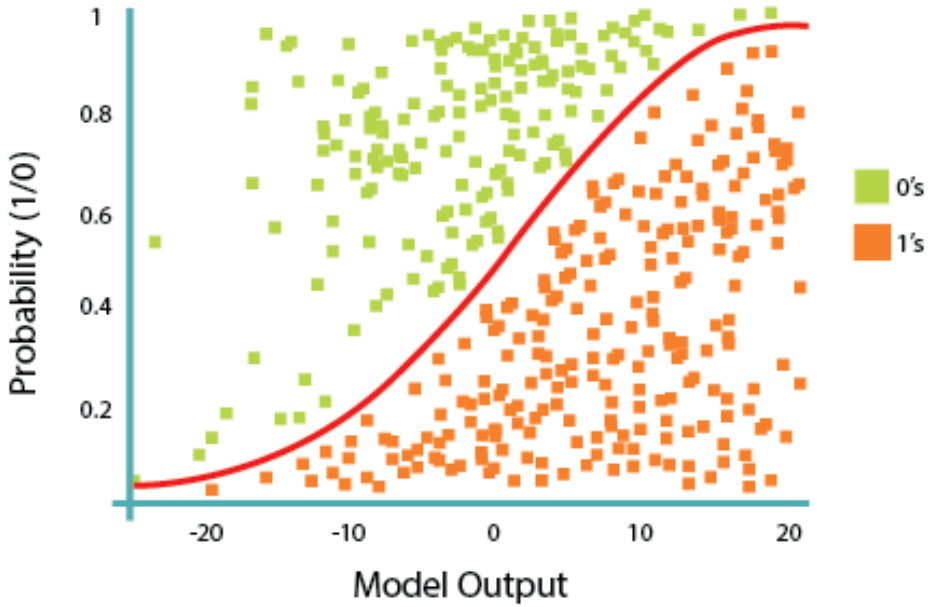
- easy to implement, interpret

Con

- constructing linear boundaries

Success Metrics:

- Confusion Matrix
- Accuracy, precision, recall, F-Score etc.

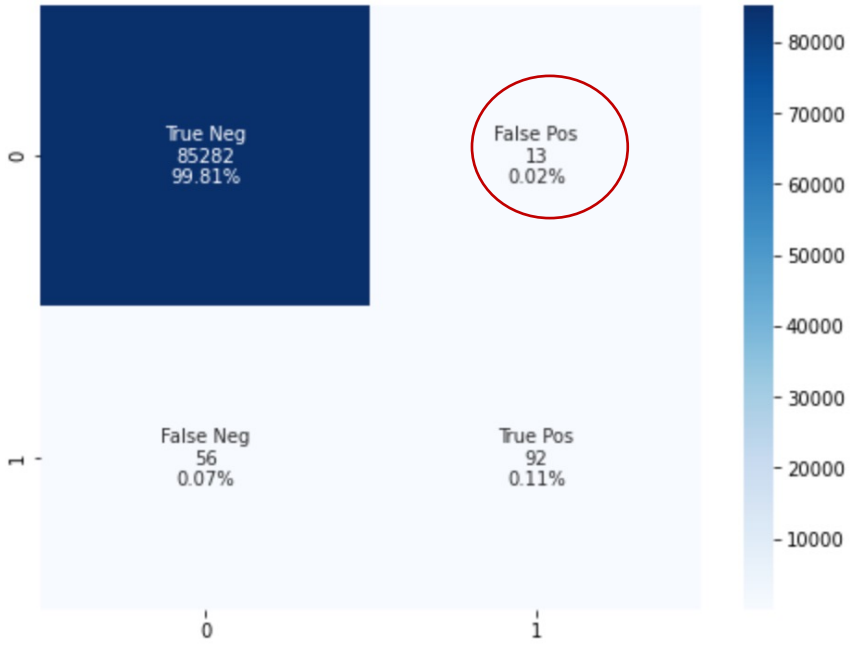


BANK OF AMERICA

Logistic Regression

Success Metrics

*** Confusion Matrix in depth analysis without SMOTE**



* Confusion Matrix for improved balance through SMOTE

BANK OF AMERICA

Logistic Regression

Success Metrics

Classification Report				
	Precision	Recall	F1 Score	Support
Non-Fraud	1.00	1.00	1.00	85295
Fraud	0.88	0.62	0.73	128
				85443
macro	0.94	0.81	0.86	85443

Interpretation

- Improved TP after SMOTE
- F1 score at 0.73
- Still a gap between our goal at 75%

Comment

- Easy to implement
- Moderate Performance
- Predictive Stage

BANK OF AMERICA

Random Forest

Supervised Machine Learning

- an ensemble of decision trees

Pro

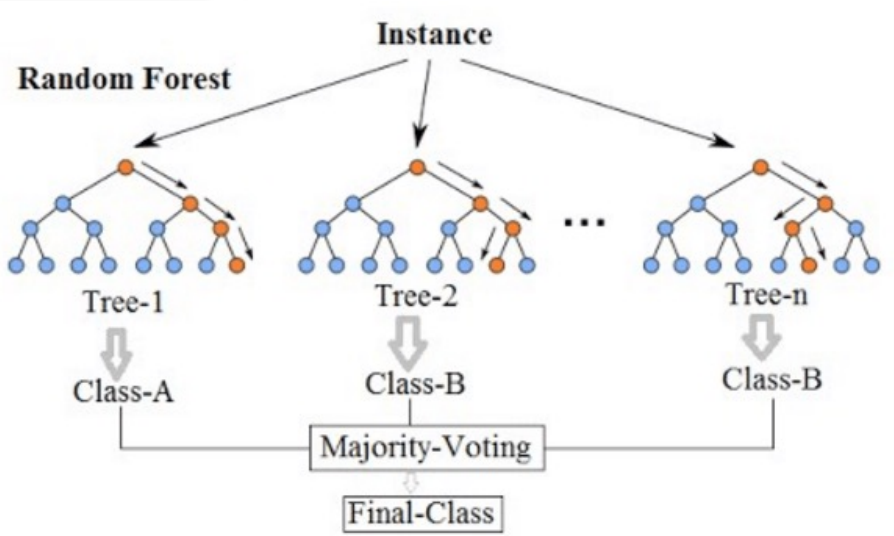
- automating missing values
- normalizing of data is not required

Con

- requiring much computational power
- time consuming

Success Metrics

- Confusion Matrix
- Classification Report with Precision, Recall, F1 score



*Confusion Matrix

BANK OF AMERICA

Random Forest

- Interpretation
  - F1 score at 79%
  - Exceed our goal at 75%

- Comment
  - Good performance
  - Open for further improvement
  - Predictive Stage

		Flagged Fraud	
		0	1
Actual Fraud	Non-Fraud	85288	7
	Fraud	46	102

Classification Report				
	Precision	Recall	F1 Score	Support
Non-Fraud	1.00	1.00	1.00	85295
Fraud	0.94	0.69	0.79	148
				85443
macro	0.97	0.84	0.90	85443

BANK OF AMERICA

Light Gradient Boosting Machine (LightGBM)

Supervised Machine Learning

- A gradient boosting framework that uses tree-based learning algorithms
- LightGBM aims to reduce complexity of histogram building (O(data * feature)) by down sampling data and feature using GOSS(Gradient Based One Side Sampling) and EFB (Exclusive Feature Bundling)
- Can be applied to **Predictive Stage**

Pro

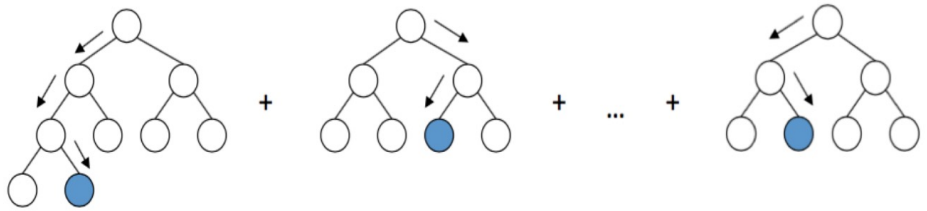
- Fast for larger dataset with better accuracy
- Weights: it balances class imbalance
- Lower memory usage

Cons

- Narrow user base due to less documentation available

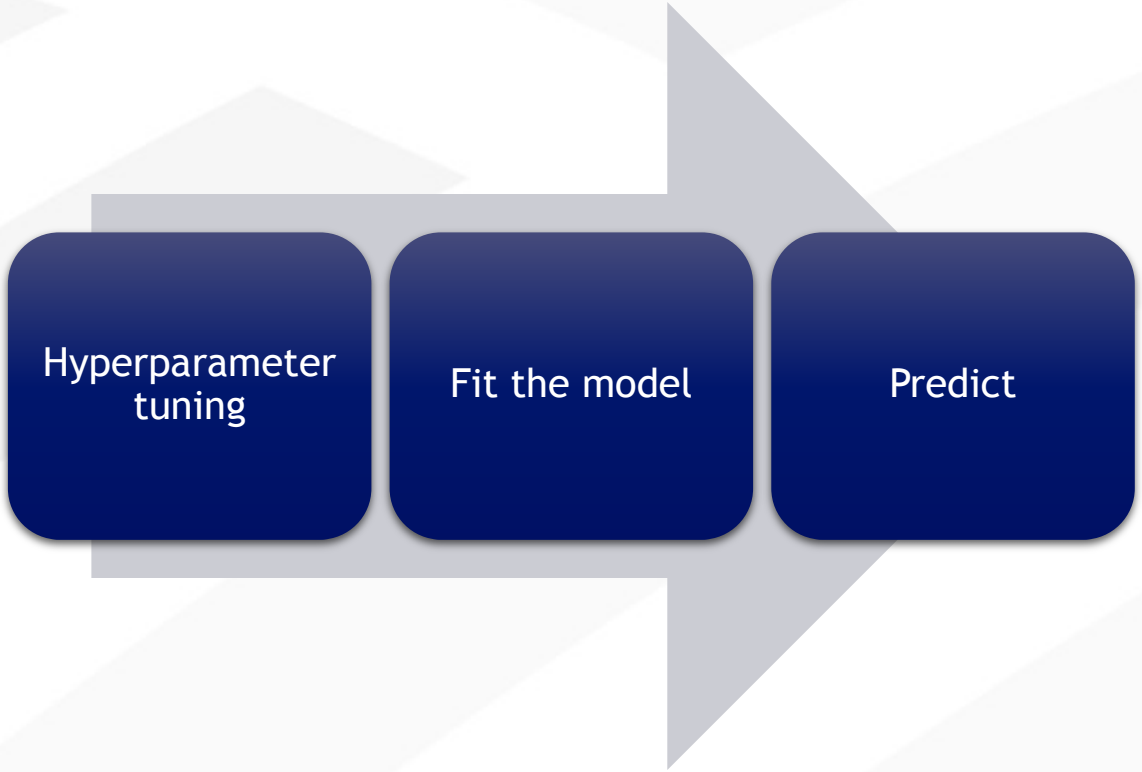
Success Metrics:

- Confusion Matrix
- Classification Report with Precision, Recall, F1 score



BANK OF AMERICA

Light Gradient Boosting Machine (LightGBM)



		Flagged Fraud	
		0	1
Actual Fraud	Non-Fraud	85255	40
	Fraud	26	122

Classification Report				
	Precision	Recall	F1 Score	Support
Non-Fraud	1.00	1.00	1.00	85295
Fraud	**0.75**	**0.82**	**0.79**	148
				85443
macro	0.87	0.92	0.89	85443

BANK OF AMERICA

Light Gradient Boosting Machine (LightGBM)

- Summary
  - Same F1 score as Random Forest
  - Fast and efficient.
  - Best model for classification so far.

Classification Report				
	Precision	Recall	F1 Score	Support
Non-Fraud	1.00	1.00	1.00	85295
Fraud	**0.73**	**0.84**	**0.78**	148
				85443
macro	0.87	0.92	0.89	85443

BANK OF AMERICA

Deep Learning Model - Keras neural network model

Supervised Deep Learning

- A neural network is a simplified model of the way the human brain processes information
- **Predictive Stage and Prescriptive Stage** (when we get more data)

Pro

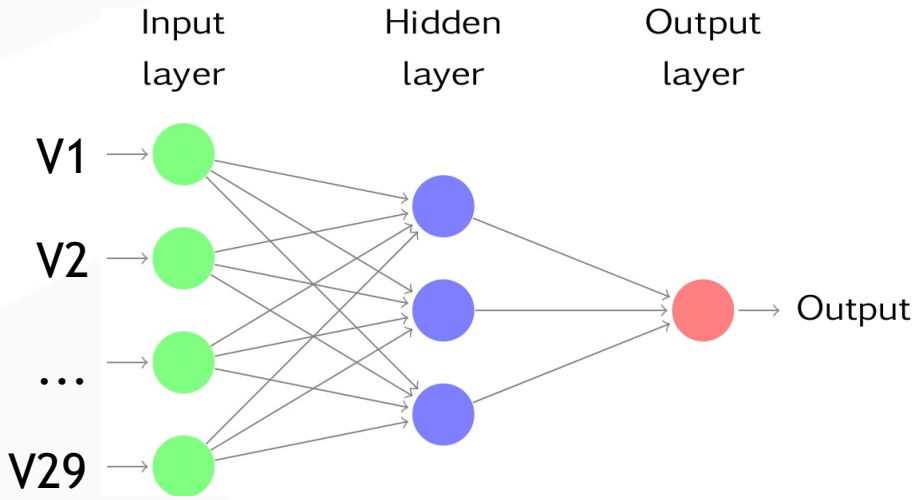
- Good to model with nonlinear data with large number of inputs
- These can be trained with any number of inputs and layers.
- Neural networks work best with more data points.

Cons

- Neural networks are black boxes, meaning we cannot know how much each independent variable is influencing the dependent variables.
- It is computationally very expensive and time consuming to train with traditional CPUs.

Success Metrics:

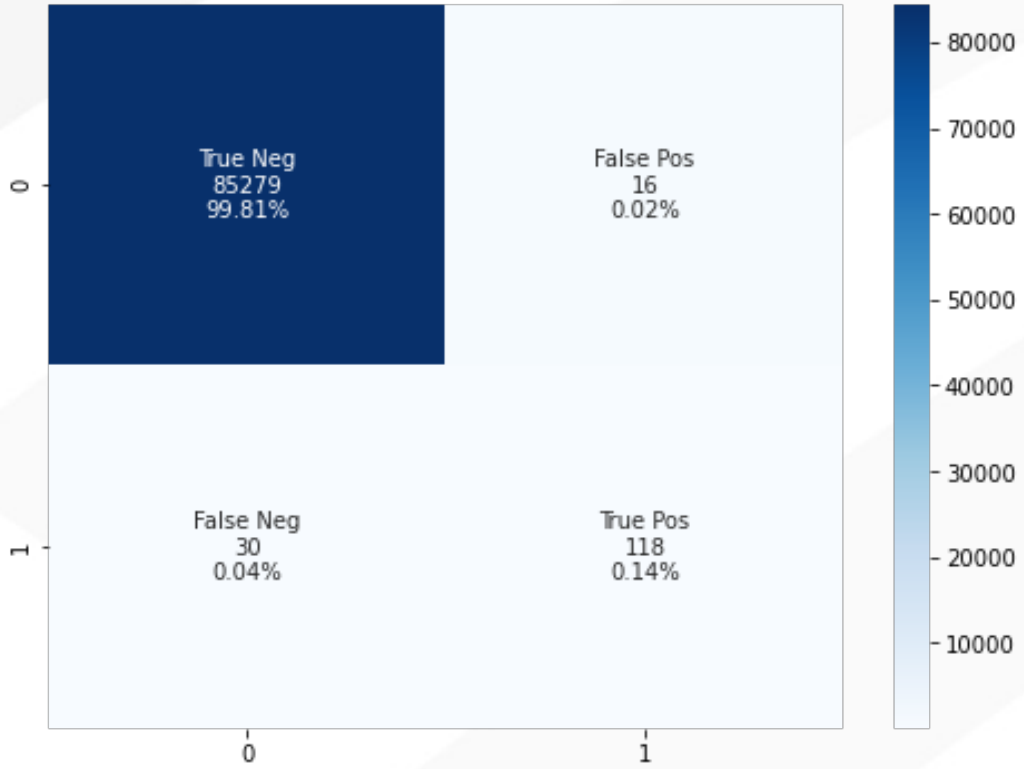
- Accuracy, F1 score, precision, recall.



BANK OF AMERICA

Keras Neural Network

Define network → Compile network → Fit network → Evaluate Network → Make precitions



Classification Report				
	Precision	Recall	F1 Score	Support
Non-Fraud	1.00	1.00	1.00	85295
Fraud	**0.89**	**0.79**	**0.84**	148
				85443

BANK OF AMERICA

Keras Neural Network

- Summary
  - Improves the accuracy further to **84%**
  - It is a deep learning model so more data means better accuracy for the model, hence as our data grows the accuracy of our model will improve further

Classification Report				
	Precision	Recall	F1 Score	Support
Non-Fraud	1.00	1.00	1.00	85295
Fraud	**0.89**	**0.79**	**0.84**	148
				85443

BANK OF AMERICA

Model Selection

	Conventional Rule-based System	Logistic Regression	Random Forest	Light GBM	Keras NN
Recall (Sensitivity)	12%	62%	69%	82%	79%
Precision	35%	88%	94%	75%	89%
F1 Score	**18%**	**73%**	**79%**	**79%**	**84%**

BANK OF AMERICA

Updated Economic Value Estimation

Best performance for deep learning case : Neural Nets

Current State

- Revenue - **$22.8B**
- Cost - **$14.4B**
- Current fraud detection algorithm - **60%**

Future State

- Increase fraud detection precision to 84%

(instead of 75%)

Gaps

- **Increase revenue**
  - Reinvesting **$3.4B** to grow market share
  - **Value $3.4B -> 6.7B**
- **Decrease costs**
  - Current Cost – **$14.4B**
  - At 83% efficiency cost will be – **$12.05B**
  - **Value $2.35B -> 3.4B**
- **Execute Both**
  - **$1.7B Save cost**

  (instead of 1.1B)
  - **$1.7B reinvest**

  (instead of 1.1B)

BANK OF AMERICA

Value Of Our Business Case

Improves both precision and recall value

+

Cancels fewer amount of non-fraud transactions

+

Detects fraud transactions with higher accuracy

=

Increasing Economic Value

Save Enormous Cost

Increase Retention Rate

Build Customer Royalty

BANK OF AMERICA

Thanks for listening!

References

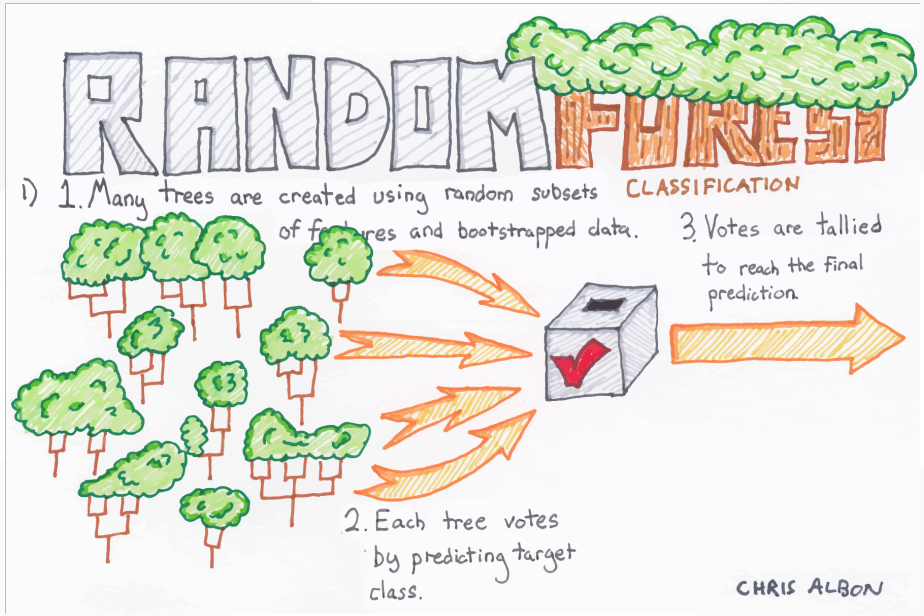
- https://www.linkedin.com/pulse/essentials-machine-learning-amit-jain
- https://towardsdatascience.com/lightgbm-vs-xgboost-which-algorithm-win-the-race-1ff7dd4917d
- https://www.tutorialspoint.com/keras/keras_deep_learning.htm
- https://www.mygreatlearning.com/blog/random-forest-algorithm/

Q&A

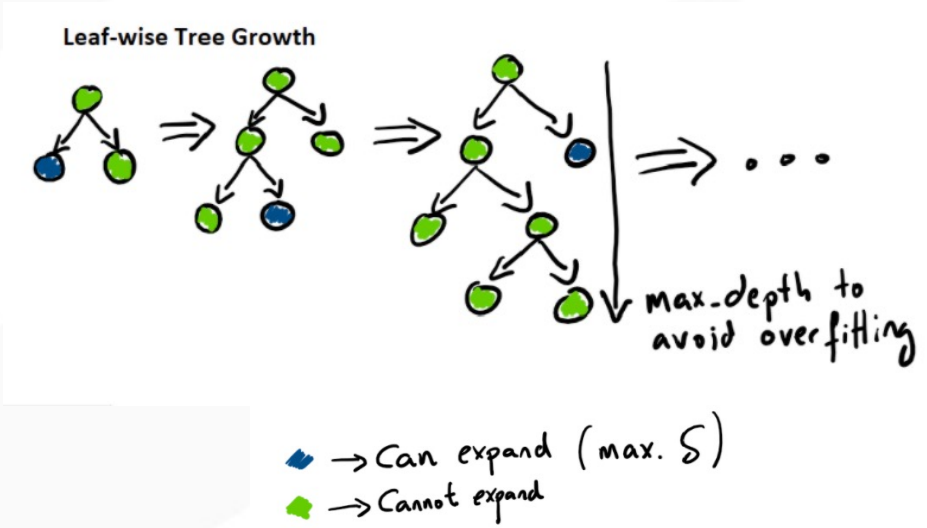
BANK OF AMERICA

Models - Decision Trees

Random Forest



Light Gradient Boosting Machine (LightGBM)



BANK OF AMERICA