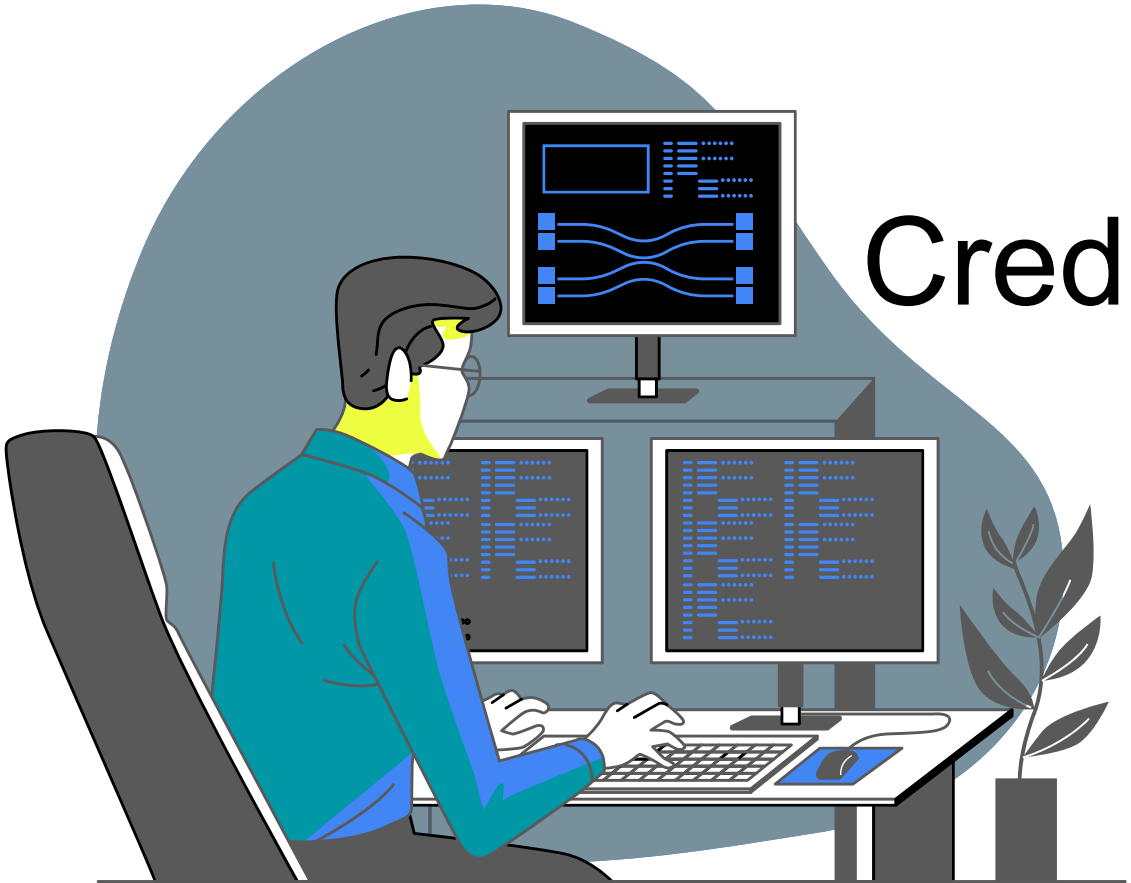


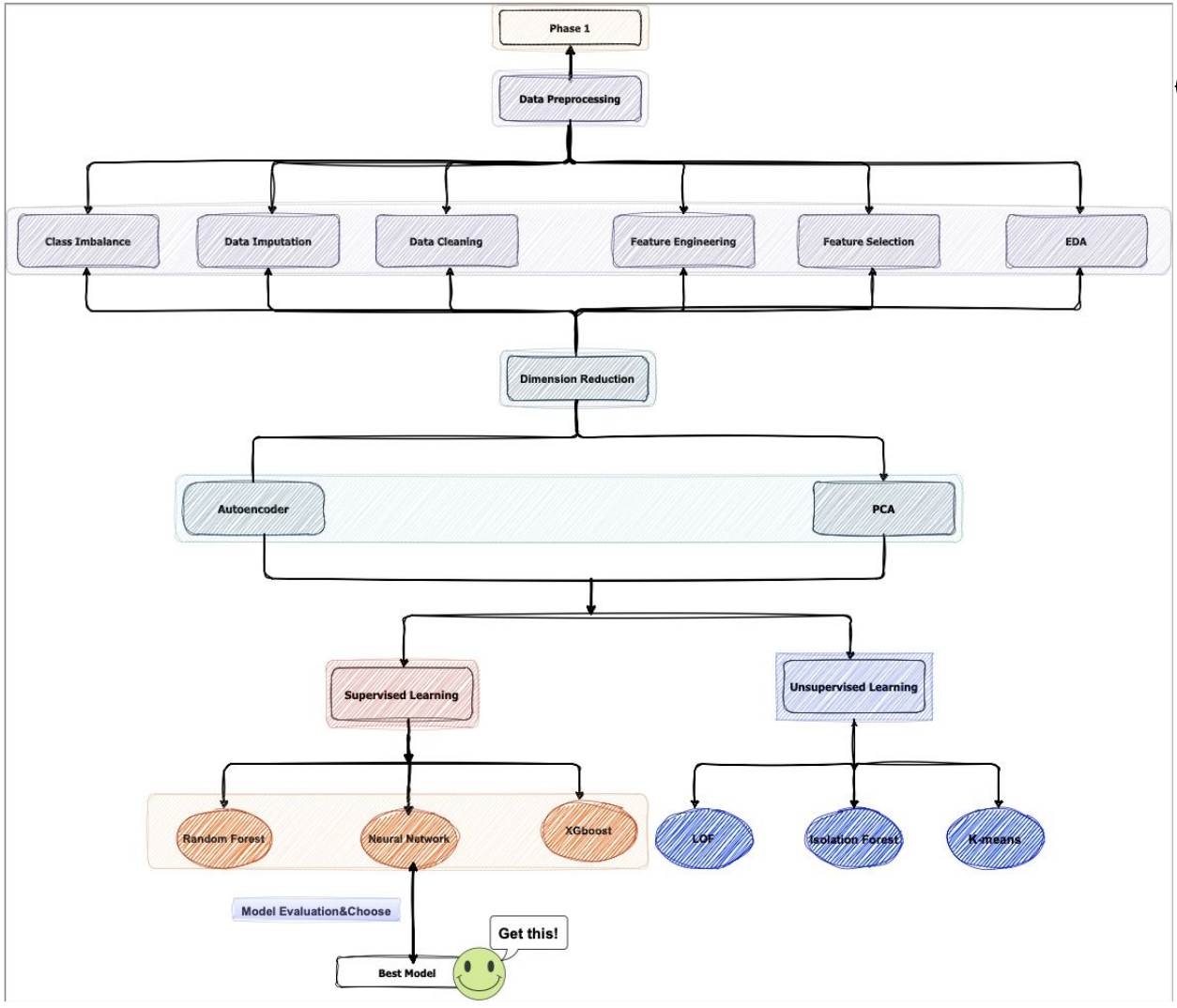
Credit Card Fraud Detection

A Hybrid Approach

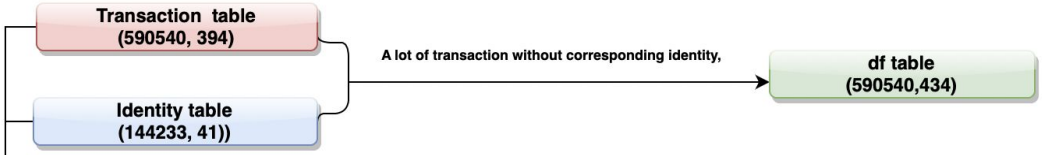
Li, Zhengyuan
Zhang, Wencheng
Zhou, Weiyan

GitHub:
https://github.com/wenchengking/dataMining

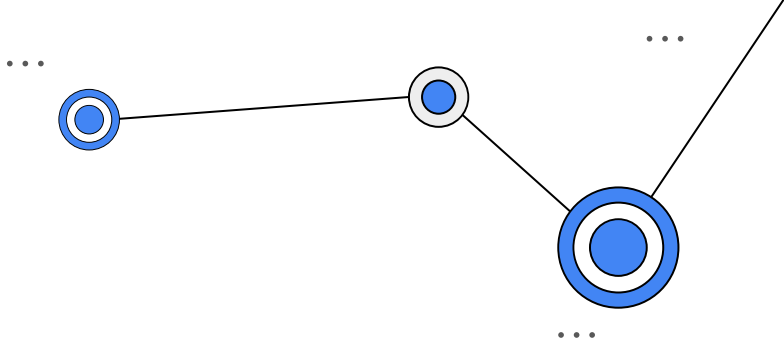
Project Pipeline



EDA



Dataset	Variable Names	Variable Description
Transaction Table	TransactionDT	timedelta from a given reference datetime (not an actual timestamp)
	TransactionAMT	transaction payment amount in USD
	ProductCD	product code, the product for each transaction
	card1 - card6	payment card information, such as card type, card category, issue bank, country, etc.
	addr	address
	dist	distance
	P_ and (R__) emaildomain	purchaser and recipient email domain
	C1-C14	counting, such as how many addresses are found to be associated with the payment card, etc. The actual meaning is masked.
	D1-D15	timedelta, such as days between previous transaction, etc.
	M1-M9	match, such as names on card and address, etc
	Vxxx	Vesta engineered rich features, including ranking, counting, and other entity relations
	isFradu	Whether this transaction is fraud or not
Identity Table	DeviceType	Variables in this table are identity information – network connection information (IP, ISP, Proxy, etc) and digital signature (UA/browser/os/version, etc) associated with transactions. The field names are masked and pairwise dictionary will not be provided for privacy protection and contract agreement
	DeviceInfo	
	id_xx	
	TransactionID	



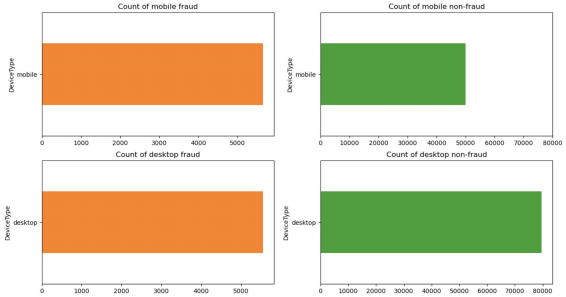
IEEE-CIS Fraud Detection

The dataset is derived from real-world e-commerce transactions conducted by Vesta, a leading provider of secure payment solutions.

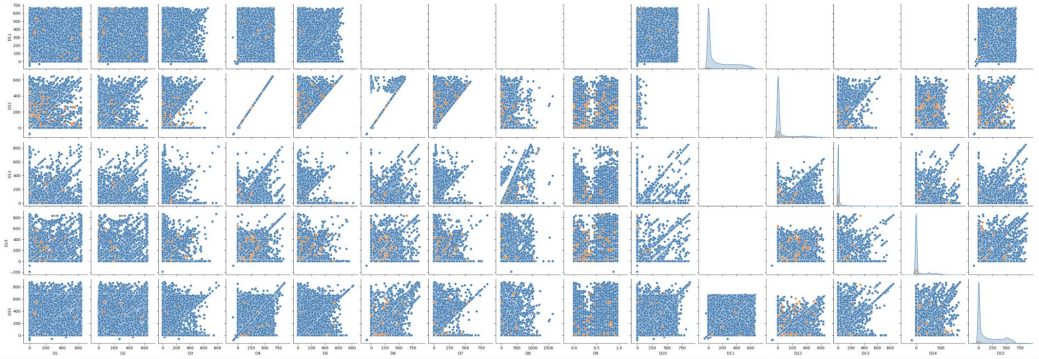
EDA

Interesting Finding

- DeviceType vs. isFraud



- D's Type ---> correlation



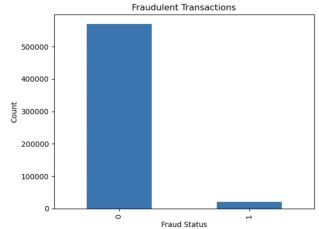
Missing values

The average number of missing value per row
195.62277407118907

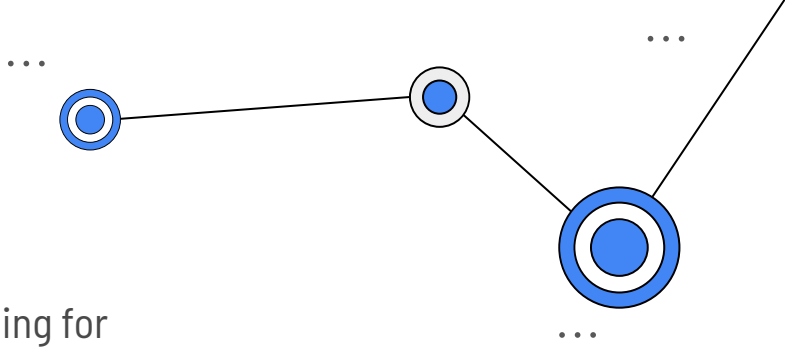
Column	% of missing value
id_24	0.991962
id_25	0.991310
id_07	0.991271
...	...
C1	0.000000
isFraud	0.000000
TransactionID	0.000000

Data imbalance

Fraud	Not fraud
569877	20663

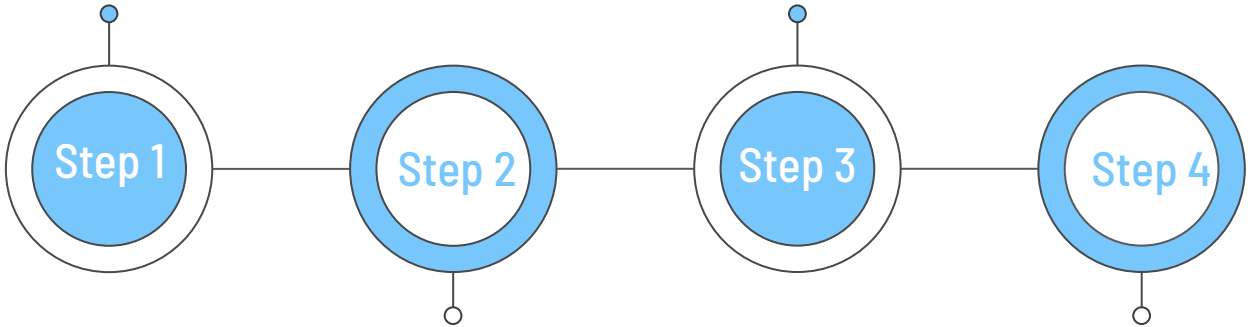


Data Engineering



Transaction Aggregation

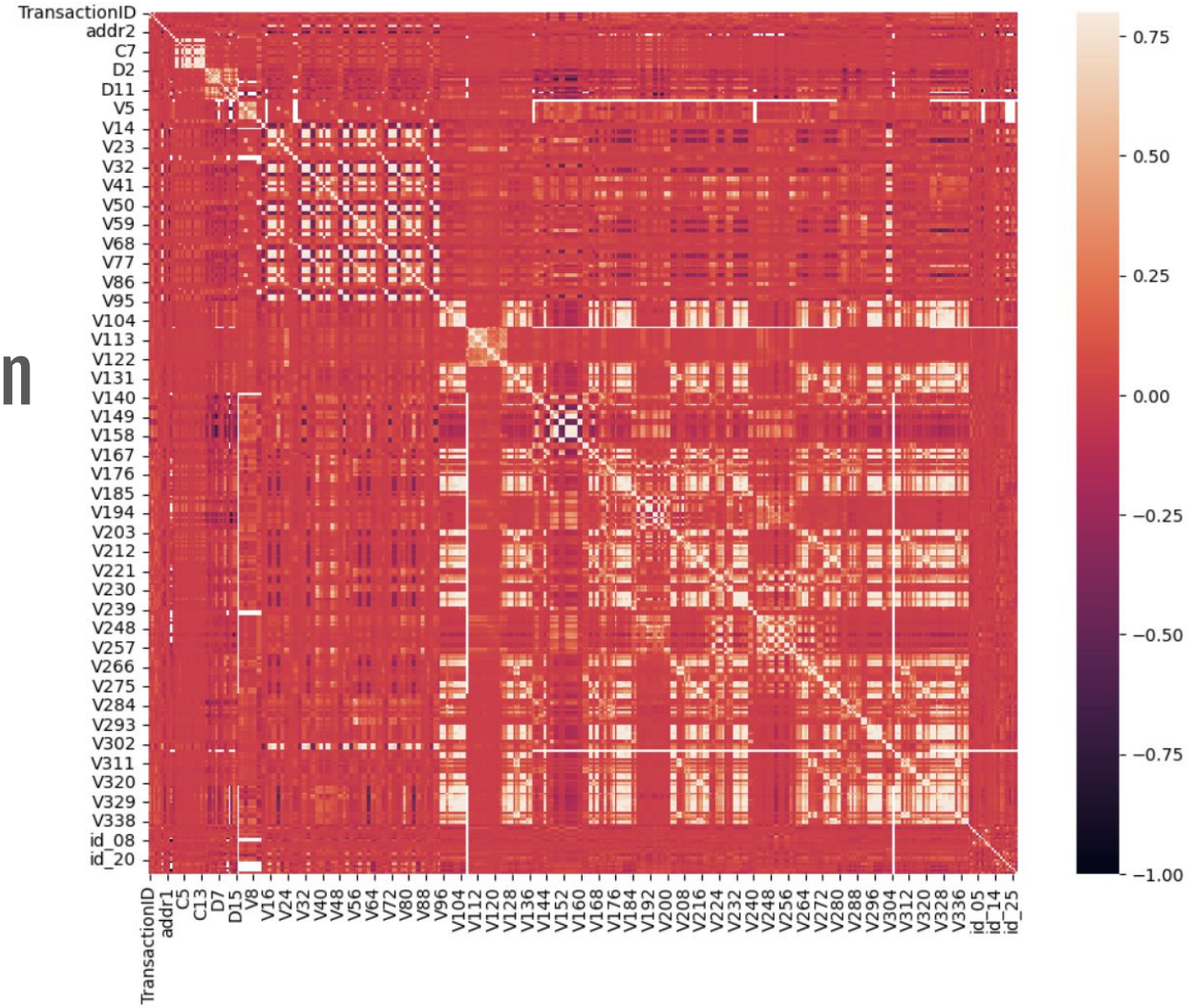
Stratified Sampling for training and testing

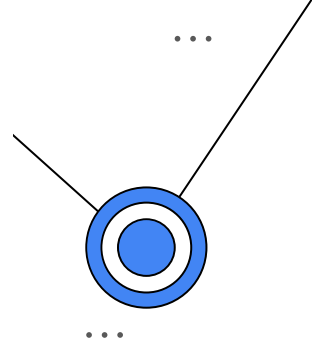


Imputing missing values via KNN imputer, and category groupings

Solving class imbalance via SMOTE

Correlation Matrix





Dimensionality Reduction

01

Linear PCA

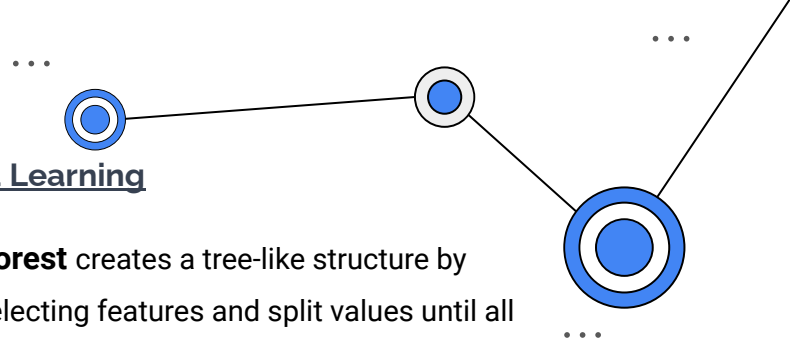
PCA is a method for reducing the complexity of data by finding its most important patterns.

02

Autoencoder

Autoencoder is a type of neural network that learns to compress and reconstruct data with high fidelity.

Model



Supervised Learning

Random Forest constructs several decision trees using a random subset of features and data, and the final output is determined by the mode of predictions from all the trees.

Neural Network comprises interconnected nodes that analyze and transmit data to forecast a class label for a given input.

XGBoost optimizes objective functions by sequentially adding decision trees to minimize the difference between predicted and actual values.

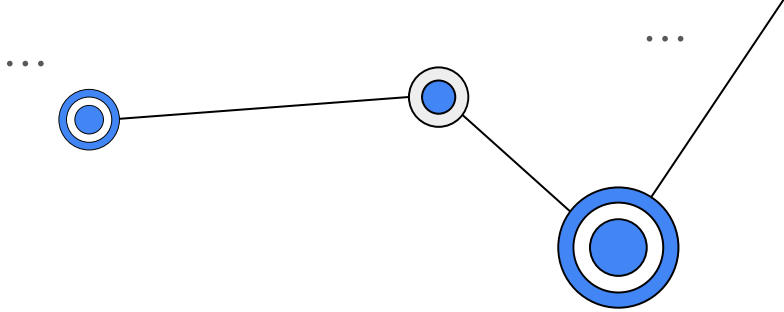
Unsupervised Learning

Isolation Forest creates a tree-like structure by randomly selecting features and split values until all instances are isolated, and the number of splits required to isolate an instance is used as an indicator of its anomaly score

LOF detect outliers by examining a data point's local deviation from its neighboring points. It does so by evaluating the density of a data point in comparison to its k nearest neighbors, flagging any data point whose density is significantly lower than that of its neighbors.

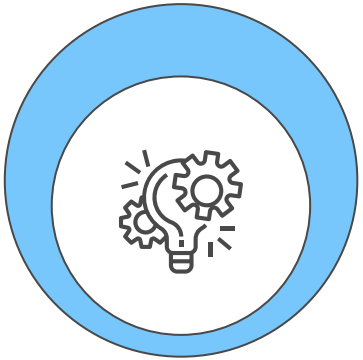
K-means groups a given dataset into a predefined number of k clusters by minimizing the sum of squared distances between each data point and its assigned cluster centroid

Model Evaluation



	CV AUC	Test AUC	Test FNR	Test FPR
★ Neural Networks Classifier	0.99998	0.99530	0.00876	0.00070
XGBoost Classifier	0.99815	0.98350	0.03220	0.00077
Random Forest Classifier	0.95881	0.87550	0.20608	0.04293
Isolation Forest	0.49300	0.50828	0.96574	0.01771
Local Outlier Factor	0.48777	0.56581	0.77331	0.09507
K Means Clustering	0.65455	0.34746	0.56826	0.73682

Improvement



Graph-Based Methods

Semi-Supervised ML

Thanks !