IEEE-CIS FRAUD DETECTION

- Ryan Timbrook
- Amanda Carvalho
- Luigi Penaloza
- Charles Cheung

Business Question and Problem to Solve

Business Question

- Improve the efficacy of fraudulent transaction alerts, helping hundreds of thousands of businesses reduce their fraud losses and increase their revenue; while securing consumer's peace of mind and wallets!

Problem to Solve

- Identify real-time fraudulent e-commerce transactions, using advanced Machine Learning algorithms, by automating alerts that block highly suspicious activities.

About the Data:

- The core data set for this project is provided by VESTA, the worlds leading payment service company, and is a kaggle competition being facilitated by the IEEE Computational Intelligence Society.
- The data is broken into two files **identity** and **transaction**, which are joined by **TransactionID**. *Not all transactions have corresponding identity information.*

Categorical Features - Transaction:

- ProductCD
- card1 card6
- addr1, addr2
- P_emaildomain
- R_emaildomain
- M1 - M9

Categorical Features - Identity:

- DeviceType
- DeviceInfo
- id_12 - 1d_38

Data Details

Transaction Table

"It contains money transfer and also other gifting goods and service, like you booked a ticket for others, etc."

- 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.

339 Features… No data definitions provided due to ILP

Categorical Features:

- ProductCD
- card1 - card6
- addr1, addr2
- Pemaildomain Remaildomain
- M1 - M9

Training dataset:

- Rows: 590, 540
- Columns: 434
- Missing Values: 414
- Total NaN count: 115,523,073

Testing dataset:

- Rows: 590, 691
- Columns: 433
- Missing Values: 414
- Total NaN count: 90,186,908

Identity Table

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. They're collected by Vesta's fraud protection system and digital security partners. (The field names are masked and pairwise dictionary will not be provided for privacy protection and contract agreement)

Categorical Features:

- DeviceType
- DeviceInfo
- id12 - id38

Data Preparation and Setup

Initial: Handling Large Data Memory Challenges

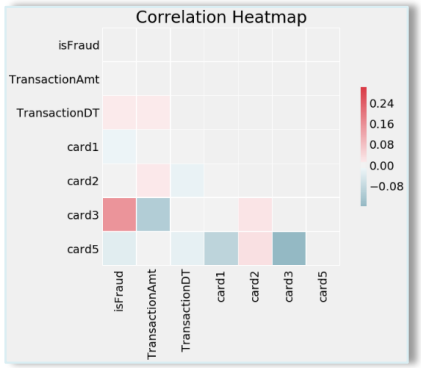
- Dataset memory reduction pre-processing steps were needed for the data to be manageable and run on google colab. On initial loading the colab session was 11GB, after running the memory reduction procedures it went down to 1.3GB.
  Training Dataset After: **memory usage: 530.0 MB**
  Testing Dataset After**: memory usage: 462.0 MB**

- At each milestone step, data objects were saved as pickle files where down stream components could load and use as needed.

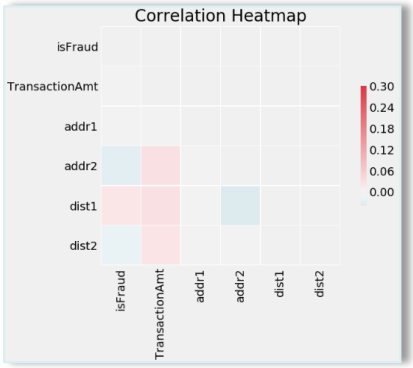
- Varying techniques in data preparation were taken for each model. The table below gives a summary.

Data Exploration
Feature Correlation & Class Imbalance

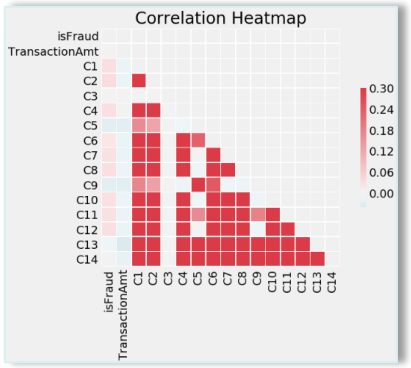
- **Card3**, Card5, TransactionDT



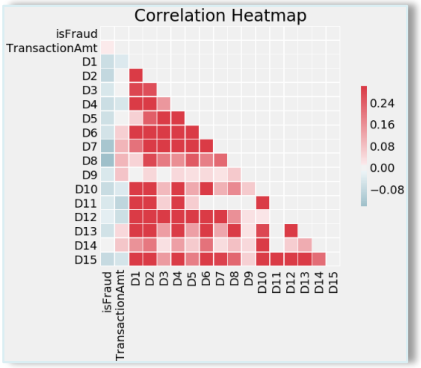
- addr2, **dest1**, dest2



- C1, C2, C6, C7, C8, C10,
- C11, C12

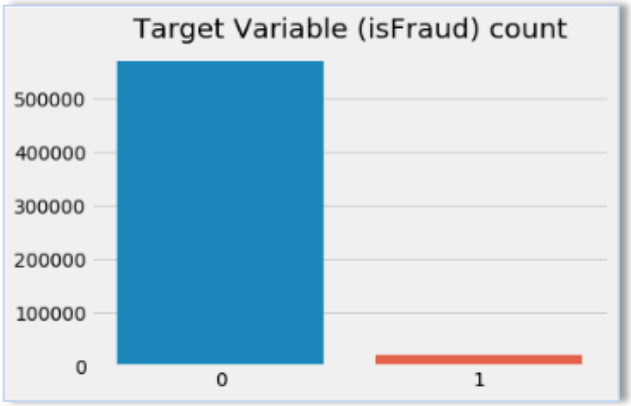


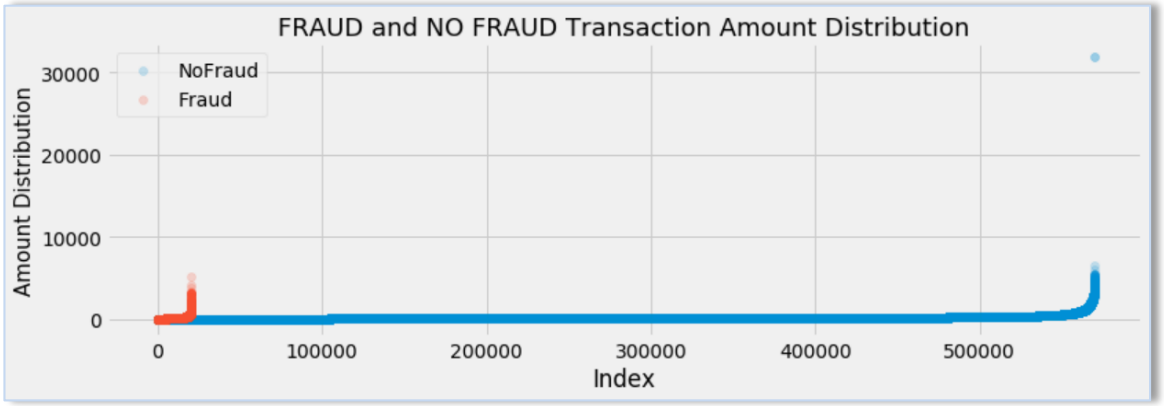
- D6, D7, **D8**, D9, D13



Representations of Class Imbalance Target Variable

- Challenges…





Descriptive Statistics

	TransactionID	isFraud	TransactionDT	TransactionAmt	card1	card2	card3	card5	addr1	addr2	dist1
count	9.282900e+04	92829.000000	9.282900e+04	92829.000000	92829.000000	92829.000000	92829.000000	92829.000000	92829.000000	92829.000000	92829.0
mean	3.167751e+06	0.047959	4.297952e+06	131.387056	9748.502203	392.048724	149.924377	198.523091	293.849153	86.751662	0.0
std	1.494700e+05	0.213681	3.997180e+06	125.589385	5000.686200	164.265423	3.376893	45.153508	99.377643	4.748481	0.0
min	2.987099e+06	0.000000	8.816200e+04	15.000000	1004.000000	0.000000	100.000000	0.000000	0.000000	0.000000	0.0
25%	3.067808e+06	0.000000	1.733386e+06	50.000000	6019.000000	264.000000	150.000000	190.000000	204.000000	87.000000	0.0
50%	3.098941e+06	0.000000	2.176484e+06	100.000000	9500.000000	399.000000	150.000000	226.000000	299.000000	87.000000	0.0
75%	3.251450e+06	0.000000	6.388409e+06	150.000000	14349.000000	553.000000	150.000000	226.000000	330.000000	87.000000	0.0
max	3.577531e+06	1.000000	1.581094e+07	1800.000000	18388.000000	600.000000	231.000000	237.000000	536.000000	102.000000	0.0

	TransactionID	isFraud	TransactionDT	TransactionAmt	ProductCD	card1	card2	card3	card4	card5	...	addr2	dist1	dist2	P_emaildomain
0	2988431	0	133636	150.0	R	11782	520	150	american express	190	...	87	0	0	anonymous.com
1	2988431	0	133636	150.0	R	11782	520	150	american express	190	...	87	0	0	anonymous.com
40	2987099	0	88162	75.0	R	1214	174	150	visa	226	...	87	0	0	gmail.com
41	2987099	0	88162	75.0	R	1214	174	150	visa	226	...	87	0	0	gmail.com
56	2987119	0	88484	100.0	H	2456	399	150	american express	118	...	87	0	0	anonymous.com

```
visa                384767
mastercard          189217
american express      8328
discover              6651
Name: card4, dtype: int64
debit               439938
credit              148986
debit or credit         30
charge card             15
Name: card6, dtype: int64
desktop     85165
mobile      55645
Name: DeviceType, dtype: int64
```

K Means Clustering

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=2, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=None, tol=0.0001, verbose=0)
```

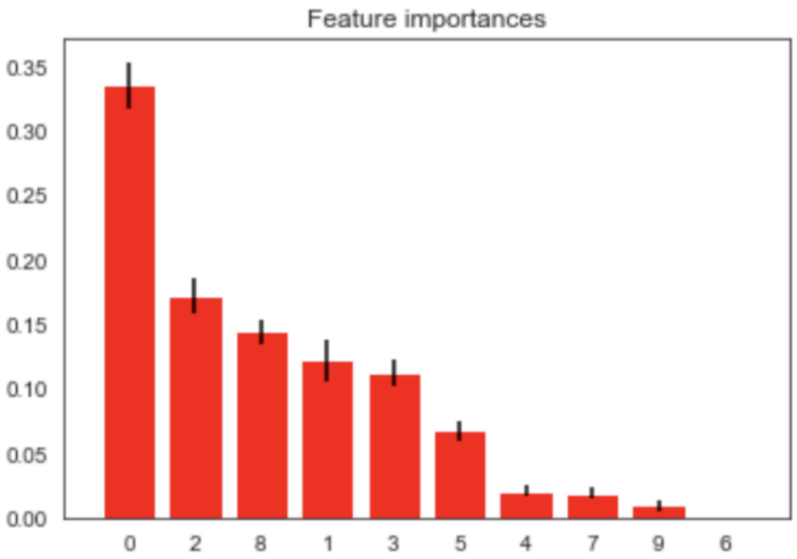
```
[52] from sklearn import preprocessing
     X = np.array(v_train_min.drop(['isFraud'], 1).astype(float))
     X = preprocessing.scale(v_train_min)
     y = np.array(v_train_min['isFraud'])

     clf = KMeans(n_clusters=2)
     clf.fit(v_train_min)

     correct = 0
     for i in range(len(X)):
         predict_me = np.array(X[i].astype(float))
         predict_me = predict_me.reshape(-1, len(predict_me))
         prediction = clf.predict(predict_me)
         if prediction[0] == y[i]:
             correct += 1

     print(correct/len(v_train_min))
```

```
0.03499000914417313
```

```
[53] X = np.array(v_train_min.drop(['isFraud', 'TransactionDT','P_emaildomain','R_emaildomain'], 1).astype(float))
     X = preprocessing.scale(v_train_min)
     y = np.array(v_train_min['isFraud'])

     clf = KMeans(n_clusters=2)
     clf.fit(v_train_min)

     correct = 0
     for i in range(len(X)):
         predict_me = np.array(X[i].astype(float))
         predict_me = predict_me.reshape(-1, len(predict_me))
         prediction = clf.predict(predict_me)
         if prediction[0] == y[i]:
             correct += 1

     print(correct/len(v_train_min))
```

```
0.9650099908558268
```

Random Forest Classifier




```
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(n_estimators = 100)
clf.fit(X_train, y_train)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=None,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```

```
forest_y_pred = clf.predict(X_test)
score = accuracy_score(y_test, forest_y_pred) * 100
rounded_score = round(score, 1)
print("Random Forest (n_est: 100) Accuracy: {}%".format(rounded_score))
```

```
Random Forest (n_est: 100) Accuracy: 98.9%
```

Features:

0 - "TransactionDT", 1 - "TransactionAmt", 2 - "card1", 3 - "card2", 4 - "card3", 5 - "card5", 6 - "dist1", 7 - "dist2", 8 - "addr1", 9 - "addr2"

Light Gradient Boosting

```
import lightgbm as lgb
```

```
for f in X_train.columns:
    if X_train[f].dtype=='object' or X_test[f].dtype=='object':
        X_train[f] = X_train[f].astype("category")
        X_train[f] = X_train[f].astype("category")
        X_test[f] = X_test[f].astype("category")
```

```
params = {'application': 'xentropy',
          'boosting': 'gbdt',
          'learning_rate': 0.1,
          'bagging_fraction': 0.6,
          'feature_fraction': 0.6,
          'verbosity': -1,
          'data_random_seed': 24,
          'early_stop': 10,
          'verbose_eval': 50,
          'num_rounds': 10000}
```

```
d_train = lgb.Dataset(X_train, label=y_train)
d_valid = lgb.Dataset(X_val, label = y_val)
watchlist = [d_train, d_valid]
num_rounds = params.pop('num_rounds')
verbose_eval = params.pop('verbose_eval')
early_stop = None
if params.get('early_stop'):
    early_stop = params.pop('early_stop')
model = lgb.train(params,
                  train_set=d_train,
                  num_boost_round=num_rounds,
                  valid_sets=watchlist,
                  verbose_eval=verbose_eval,
                  early_stopping_rounds=early_stop)
```

```
lgb_pred_val = model.predict(X_val, num_iteration=model.best_iteration)
print(roc_auc_score(y_val, lgb_pred_val))
```

```
0.9443788960119005
```

Base Neural Network
Issues with Class Imbalance

Base Model Hyperparameters:

- Kernel_initializer: normal
- Dense Input Layer:
- Activation Function: **'relu'**
- Dense Output Layer:
- Activation Function: **'softmax'**
- Compiler:
- loss: **'categorical_crossentropy**
- optimizer: **'adam'**
- Built on:
- epochs: **300**
- batch_size: **1000**

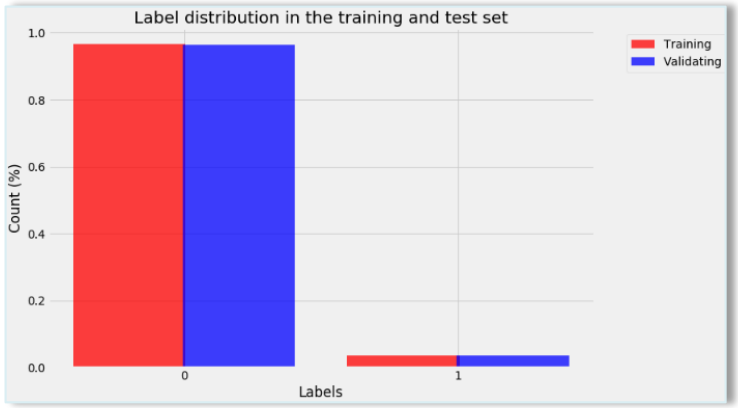
Model Summary:

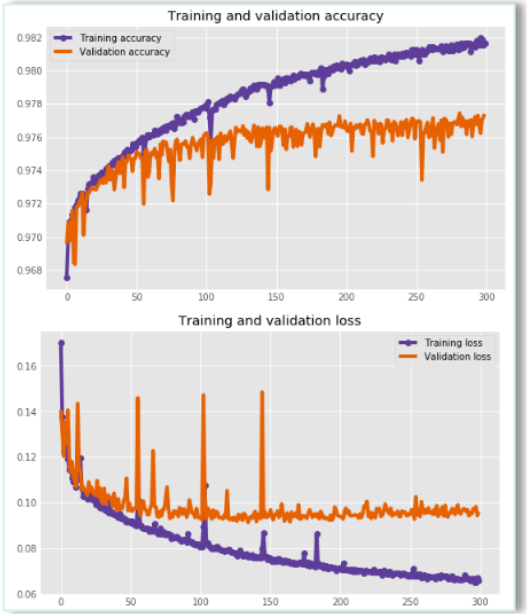
```
None
254520
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 503)               253512
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 1008
=================================================================
Total params: 254,520
Trainable params: 254,520
Non-trainable params: 0
_________________________________________________________________
```

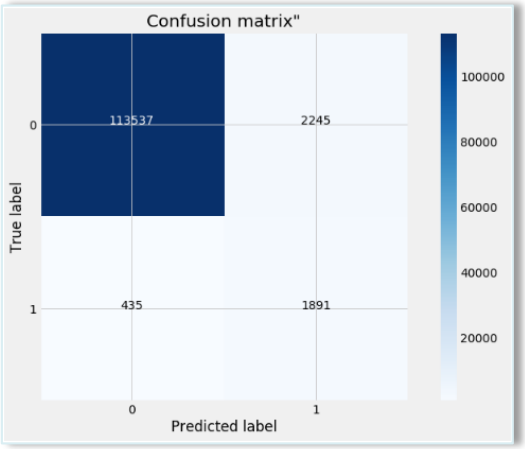
```
Test loss: 0.0959736775085113
Test accuracy: 0.9773089037192809
Baseline Error: 2.2691096280719023
```

	precision	recall	f1-score	support
Class0	0.98	1.00	0.99	113972
Class1	0.81	0.46	0.59	4136
accuracy			0.98	118108
macro avg	0.90	0.73	0.79	118108
weighted avg	0.97	0.98	0.97	118108

```
X_train shape: (330702, 205)
y_train shape: (330702, 2)
X_test shape: (118108, 205)
y_test shape: (118108, 2)
X_val shape: (141730, 205)
y_val shape: (141730, 2)
```







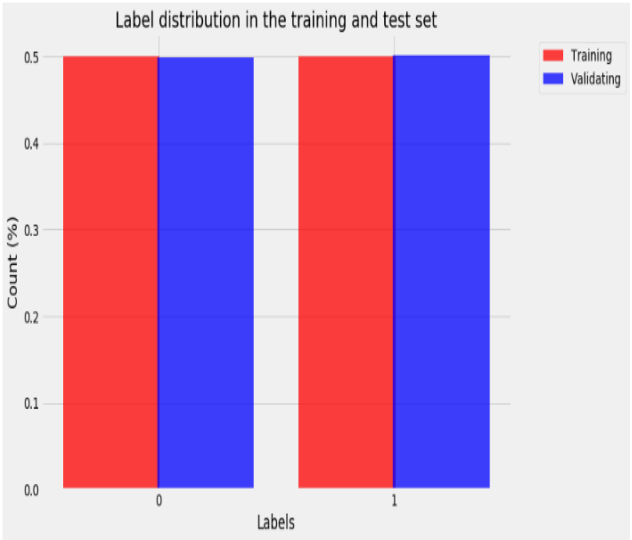
Base Neural Network
Issues with Class Imbalance and Upsampling

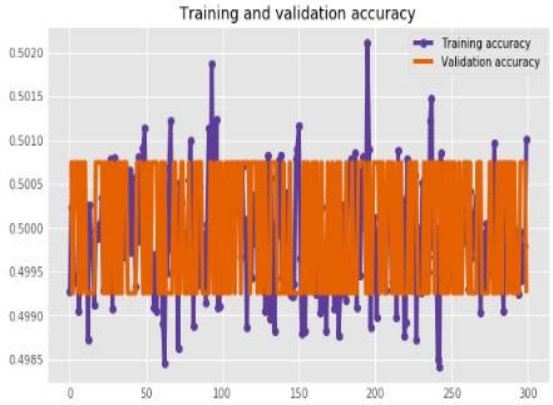
```
None
254520
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_1 (Dense)              (None, 503)               253512
_________________________________________________________________
dense_2 (Dense)              (None, 2)                 1008
=================================================================
Total params: 254,520
Trainable params: 254,520
Non-trainable params: 0
_________________________________________________________________
```

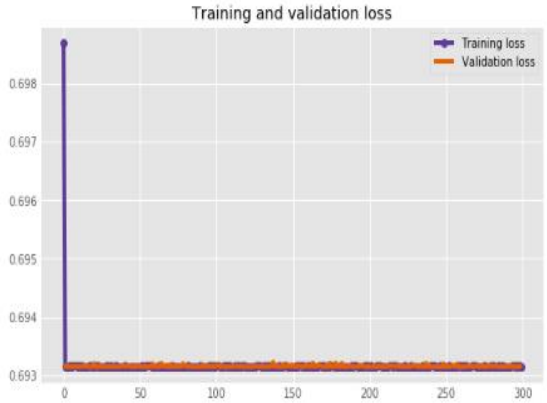
Test loss: 0.6859523552970881 Test accuracy: 0.9649812036380664
Baseline Error: 3.5018796361933653

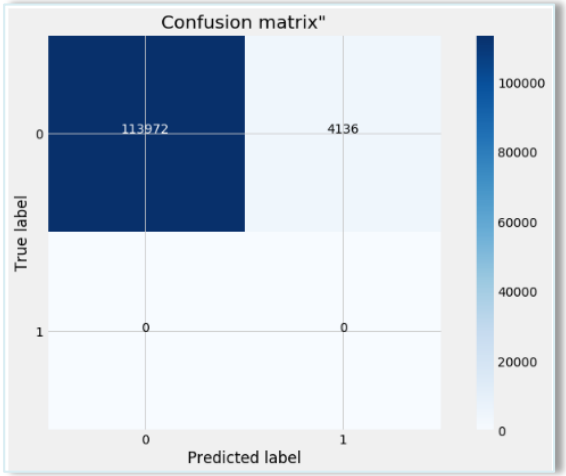
	precision	recall	f1-score	support
Class0	0.96	1.00	0.98	113972
Class1	0.00	0.00	0.00	4136
accuracy			0.96	118108
macro avg	0.48	0.50	0.49	118108
weighted avg	0.93	0.96	0.95	118108

Original data split, Not Fraud: (455905, 504)
Original data split, Is Fraud: (16527, 504)
Up sampled isFraud count: 1.0 455905 0.0 455905









Results and Recommendations

Model	Model Accuracy
Random Forest Classifier	**98.9%**
K Means Clustering	**96.5%**
Light Gradient Boosting	**94.4%**
Base Neural Network	**98%**

Questions?

Data Definitions

Identity Data Attribute Definitions

Column	Description	Categorical or Numerical
DeviceType	Mobile or Desktop	Categorical
DeviceInfo	Brand, model, make, build of device	Categorical
id01 - id11	-	Numerical
id12	Found/NotFound	Categorical
id13	-	Categorical
id14	-	Categorical
id15	Found/NotFound/Unknown	Categorical
id16	Found/NotFound	Categorical
id17	-	Categorical
id18	-	Categorical
id19	-	Categorical
id20	-	Categorical
id21	-	Categorical
id22	-	Categorical
id23	IP proxy	Categorical
id24	-	Categorical
id25	-	Categorical
id26	-	Categorical
id27	Found/NotFound	Categorical
id28	Found/New	Categorical
id29	Found/NotFound	Categorical
id30	OS version	Categorical
id31	Brower version	Categorical
id32	0/16/24/32	Categorical
id33	(possibly) screen size	Categorical
id34	Matchstatus: -1/0/1/2	Categorical
id35	T/F	Categorical
id36	T/F	Categorical
id37	T/F	Categorical
id38	T/F	Categorical

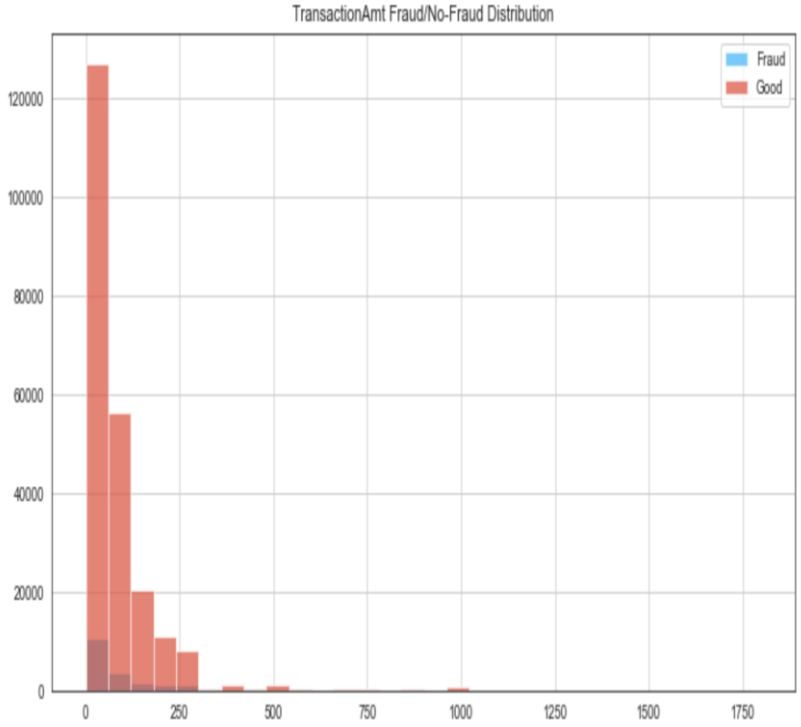
Transaction Data Attribute Definitions

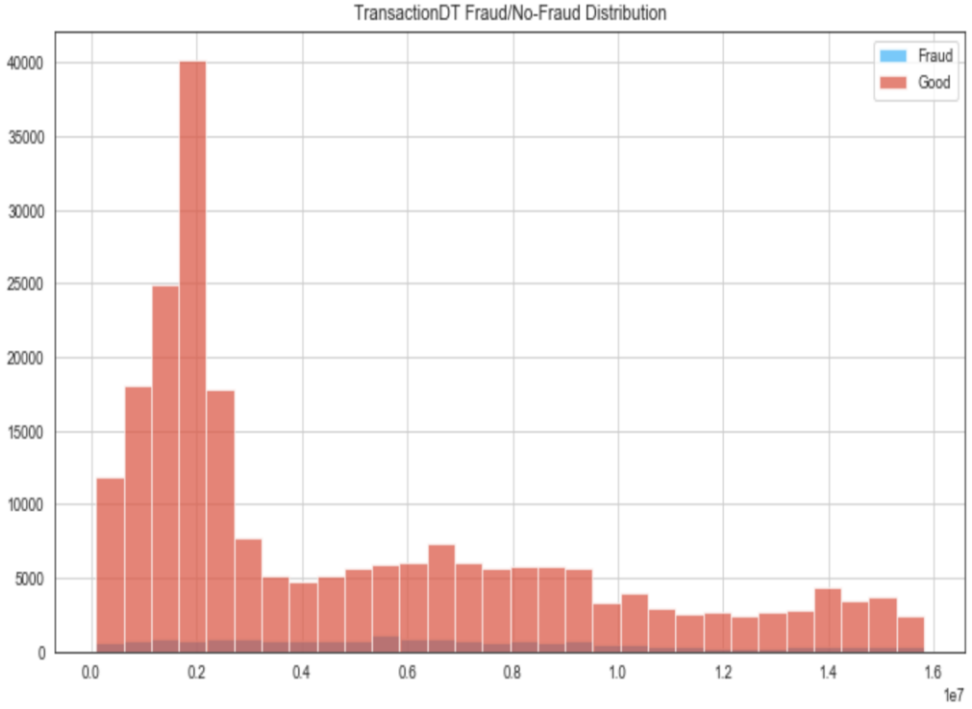
Column	Description	Categorical or Numerical
TransactionID	Transaction index	Numerical
isFraud	•Logic of labeling: reported chargeback on the card as fraud transactions and transactions posterior to it with either user account, email address or billing address directly linked to these attributes as fraud too. •If none of above is reported and found beyond 120 days, then legit transaction (isFraud=0).	Categorical
TransactionDT	•timedelta from a given reference datetime (not an actual timestamp) •time index •test set transactions occurred at a later timeframe (relative to train set)	Numerical
TransactionAmt	•Logic of labeling: reported chargeback on the card as fraud transactions and transactions posterior to it with either user account, email address or billing address directly linked to these attributes as fraud too. •If none of above is reported and found beyond 120 days, then legit transaction (isFraud=0).	Numerical
ProductCD	•product code, the product for each transaction •5 unique categories	Categorical
card1	payment card information, such as card type, card category, issue bank, country, etc.	Categorical
card2	payment card information, such as card type, card category, issue bank, country, etc.	Categorical
card3	payment card information, such as card type, card category, issue bank, country, etc.	Categorical
card4	Issuer (e.g. amex, visa)	Categorical
card5	payment card information, such as card type, card category, issue bank, country, etc.	Categorical
card6	type of card (e.g. debit, credit)	Categorical
addr1	purchaser billing region	Categorical
addr2	purchaser country	Categorical
P_emaildomain	purchaser email domain	Categorical
R_email_domain	recipient email domain	Categorical
M1 - M9	Numerical	Categorical
dist1	distances between (not limited) billing address, mailing address, zip code, IP address, phone	Categorical
dist2	distances between (not limited) billing address, mailing address, zip code, IP address, phone	Categorical
C1 - C14	counting, such as how many addresses are found to be associated with the payment card, etc.	Numerical
D1 - D15	timedelta, such as days between previous transaction, etc.	Numerical
V1 - V339	Vesta engineered rich features, including ranking, counting, and other entity relations.	Numerical

Appendix: Transaction Data Definitions

Column	Description	Categorical or Numerical
TransactionID	Transaction index	Numerical
isFraud	•Logic of labeling: reported chargeback on the card as fraud transactions and transactions posterior to it with either user account, email address or billing address directly linked to these attributes as fraud too. •If none of above is reported and found beyond 120 days, then legit transaction (isFraud=0). •Mislabelling (e.g. due to unreported cases) can be considered to be "unusual cases and negligible portion"	Categorical
TransactionDT	•timedelta from a given reference datetime (not an actual timestamp) •time index •test set transactions occurred at a later timeframe (relative to train set)	Numerical
TransactionAmt	•Logic of labeling: reported chargeback on the card as fraud transactions and transactions posterior to it with either user account, email address or billing address directly linked to these attributes as fraud too. •If none of above is reported and found beyond 120 days, then legit transaction (isFraud=0). •Mislabelling (e.g. due to unreported cases) can be considered to be "unusual cases and negligible portion"	Numerical
ProductCD	•product code, the product for each transaction •5 unique categories	Categorical
card1	payment card information, such as card type, card category, issue bank, country, etc.	Categorical
card2	payment card information, such as card type, card category, issue bank, country, etc.	Categorical
card3	payment card information, such as card type, card category, issue bank, country, etc.	Categorical
card4	Issuer (e.g. amex, visa)	Categorical
card5	payment card information, such as card type, card category, issue bank, country, etc.	Categorical
card6	type of card (e.g. debit, credit)	Categorical
addr1	purchaser billing region	Categorical
addr2	purchaser country	Categorical
P_emaildomain	purchaser email domain	Categorical
R_email_domain	recipient email domain	Categorical
M1 - M9	Numerical	Categorical
dist1	distances between (not limited) billing address, mailing address, zip code, IP address, phone area, etc.	Categorical
dist2	distances between (not limited) billing address, mailing address, zip code, IP address, phone area, etc.	Categorical
C1 - C14	counting, such as how many addresses are found to be associated with the payment card, etc. The actual meaning is masked.	Numerical
D1 - D15	timedelta, such as days between previous transaction, etc.	Numerical
V1 - V339	Vesta engineered rich features, including ranking, counting, and other entity relations.	Numerical

Transaction Fraud/No Fraud Credit Card Distribution





Data Preparation and Setup

Initial: Handling Large Data Memory Challenges

- Dataset memory reduction pre-processing steps were needed for the data to be manageable and run on google colab. On initial loading the colab session was 11GB, after running the memory reduction procedures it went down to 1.3GB.
  Training Dataset After: **memory usage: 530.0 MB**
  Testing Dataset After**: memory usage: 462.0 MB**

- At each milestone step, data objects were saved as pickle files where down stream components could load and use as needed.

- Varying techniques in data preparation were taken for each model. The table below gives a summary.

Model	Cleaning	Transformation	Normalization	Data Shape After Transformations
Neural Network	Narrowed Feature set to Transaction Table attributes - 64 total - Continuous - Numeric	Each of the features were individually cleaned and transformed based on it's datatype and characteristics. Effort was to retain as much of the original dataset as possible.	One Hot encoding of categorical and class features.	X shape: (472432, 503) y shape: (472432,) X_test shape: (118108, 503) y_test shape: (118108,) test_comp shape: (506691, 503)