CREDIT CARD TRANSACTION FRAUD REPORT

TABLE OF CONTENTS

TABLE OF FIGURES:

EXECUTIVE SUMMARY:

The following report aims to help understand the better model to understand to choose for better savings from the happening of fraud.
The report starts with a brief introduction of the dataset, the statistics, and field summary tables. Data cleaning describes the process of how each field is cleaned.
It follows a series of steps that are elaborated on below and as a threshold, a Fraud Detection Rate (FDR) of 3% is used. The report also provides reasoning for choosing the below model for the analysis.
The data is split into training and testing, and validation (out of time (oot – considering the last 2 months of data not involved in training or testing))
The brief conclusions and summary are provided, with an appendix that includes the data quality report visualizations.
The names mentioned in double quotations ("..") indicate the field names.
The appendix

METHODOLOGY:

INTRODUCTION:

The dataset "card transactions data" corresponds to Transaction Fraud data, which concentrates on the Personal Identity Information (PII) of the individual when the card is being used. This data is synthetic and is generated in a way to mimic the real data. The company considered is based in Tennessee. This data covers a period of 365 days in the year 2010, with 10(18 - 8 unclean) fields and 96,753 records.

STATISTICS TABLES:

a) Before changing to respective data types:

	Recnum	Cardnum	Date	Merchnum	Merch description	Merch state	Merch zip	Transtype	Amount	Fraud
count	96753.000000	9.675300e+04	96753	93378	96753	95558	92097.000000	96753	9.675300e+04	96753.000000
unique	NaN	NaN	365	13091	13126	227	NaN	4	NaN	NaN
top	NaN	NaN	2/28/10	930090121224	GSA-FSS-ADV	TN	NaN	P	NaN	NaN
freq	NaN	NaN	684	9310	1688	12035	NaN	96398	NaN	NaN
mean	48377.000000	5.142202e+09	NaN	NaN	NaN	NaN	44706.596740	NaN	4.278857e+02	0.010945
std	27930.329635	5.567084e+04	NaN	NaN	NaN	NaN	28369.537945	NaN	1.000614e+04	0.104047
min	1.000000	5.142110e+09	NaN	NaN	NaN	NaN	1.000000	NaN	1.000000e-02	0.000000
25%	24189.000000	5.142152e+09	NaN	NaN	NaN	NaN	20855.000000	NaN	3.348000e+01	0.000000
50%	48377.000000	5.142196e+09	NaN	NaN	NaN	NaN	38118.000000	NaN	1.379800e+02	0.000000
75%	72565.000000	5.142246e+09	NaN	NaN	NaN	NaN	63103.000000	NaN	4.282000e+02	0.000000
max	96753.000000	5.142847e+09	NaN	NaN	NaN	NaN	99999.000000	NaN	3.102046e+06	1.000000

Figure 1 :Statistics before changing the data type

b) After changing to respective data types:

	Recnum	Cardnum	Merch zip	Amount	Fraud
count	96753.000000	9.675300e+04	92097.000000	9.675300e+04	96753.000000
mean	48377.000000	5.142202e+09	44706.596740	4.278857e+02	0.010945
std	27930.329635	5.567084e+04	28369.537945	1.000614e+04	0.104047
min	1.000000	5.142110e+09	1.000000	1.000000e-02	0.000000
25%	24189.000000	5.142152e+09	20855.000000	3.348000e+01	0.000000
50%	48377.000000	5.142196e+09	38118.000000	1.379800e+02	0.000000
75%	72565.000000	5.142246e+09	63103.000000	4.282000e+02	0.000000
max	96753.000000	5.142847e+09	99999.000000	3.102046e+06	1.000000

Figure 2 : Statistics post changing the data types

FIELD SUMMARY TABLES:

a) Numeric Fields:

Field Name	% Populated	Min	Max	Mean	Stdev	%Zero
Date	100%	2010-01-01	2010-12-31	-	-	0
Amount	100%	0.01	3102045.53	-	-	0

Figure 3 : Numeric fields summary

b) Categorical Fields:

Field Name	% Populated	Unique Values	Most Common Values
Recnum	100%	96,753	NA
Cardnum	100%	1,645	5142148452
Merchnum	96.51%	13,091	930090121224
Merch Description	100%	13,125	GSA-FSS-ADV
Merch state	98.76%	227	TN
Merch zip	95.18%	4,567	38118
Transtype	100%	4	P
Fraud	100%	2	0

Figure 4 : Categorical Fields Summary

DATA CLEANING:

FIELD IMPUTATION:

1. Null values are recognized in Merch num, Merch state, and Merch Zip.
2. Firstly, all transaction types other than 'P' and Amount > 3000000 are removed.

MERCHNUM:

3. The "Merchnum" column is filled first by creating a dictionary of not null values between the "Merch description" and "merchnum." This dictionary is used to map to the "merchnum" field to fill NA's.
4. A merchnum create a dictionary is created to store new merchnum values for each unique merch description value. These values are made by adding +1 to every previously calculated max merchnum value to ensure the values are unique.

MERCH STATE:

5. The "Merch state" field is filled in the same manner by creating a dictionary of unique values with null "merch state" values and not null "merch zip" values, which are mapped to the merch state to fill in the merch state.
6. Another two dictionaries between not null "merchnum" and "merch state," not null "merch description," and "merch state" is created and mapped to "merch state" to fill in the remaining values.
7. The remaining null values are filled by manually inputting the state codes.

MERCH ZIP:

8. "Merch zip" is filled by creating dictionaries with not null "merchnum" and "merch zip," not null "merch description," and "merch zip" to map to the "merch zip" column.
9. The remaining NA's are filled by an 'unknown' placeholder value.
10. The values in the "Merch description" for "Retail Credit Adjustment" and "Retail Debit Adjustment" are masked to "unknown" placeholder values for "merchnum," "merch state," and "merch zip."

VARIABLE GENERATION:

Variables made – Day since variables, Frequency & Amount Variables, degree centrality network variables, velocity change variables, velocity variables, velocity days since ratio, max indicator variables, acceleration variables, variability.

I have included a new variable which is a part of Network Variables known as *Degree Centrality*, which measures the # of connections of each entity in the transaction network with other entities. A high degree of centrality could indicate a central role in a fraudulent network.

Table 1 : Generated Variables Count

DESCRIPTION OF VARIABLES	**# VARIABLES CREATED**
ORIGINAL FIELDS FROM THE DATASET EXCLUDING 'RECORD' AND 'FRAUD'	(ALREADY EXISTING: 8)
TARGET VARIABLE('fraud')	(ALREADY EXISTING: 1)
RECORD VARIABLE('record')	(ALREADY EXISTING: 1)
CHECK_DATE COLUMN (TO CREATE DAY SINCE VARIABLES SELECTING THE LAST(MOST RECENT) 'DATE' AND 'CHECK_DATE')	1
CHECK_RECORD COLUMN (TO CREATE DAY SINCE VARIABLES WHERE 'RECNUM' IS GREATER THAN 'CHECK_RECORD')	1
DATE/DAY OF WEEK TARGET ENCODED ('DOW)	1
DATE/DAY OF WEEK RISK PERCENTAGE (AVERAGE FRAUD PERCENTAGE OF THAT DAY) ('DOW_RISK')	1
MERCHANT STATE TARGET ENCODED(MS_RISK)	1
NEW ENTITIES COMBINING/CONCATENATING DIFFERENT ORIGINAL FIELDS	13
DAYS SINCE VARIABLES AND FREQUENCY AND AMOUNT VARIABLES : DAYS SINCE VARIABLES: # OF DAYS SINCE THAT ENTITY/ATTRIBUTE WAS LAST SEEN FREQUENCY VARIABLES : # OF TRANSACTIONS OVER {0,1,3,7,14,30,60} AMOUNT VARIABLES : AMOUNT IN EACH TRANSACTION (CONSIDERED AGG : MEAN, MEDIAN, MAX, TOTAL, ACTUAL)	1153
DEGREE CENTRALITY NETWORK VARIABLES	18
VELOCITY CHANGE VARIABLES	180
VELOCITY DAYS SINCE RATIO VARIABLES	36
MAXIMUM INDICATOR VARIABLES:THESE VARIABLES REFER TO THE MAXIMUM COUNT OF AN ENTITY, GROUPED BY OTHER ENTITIES FOR {1, 3, 7, 30} DAYS	90
ACCELERATION VARIABLES	144
VARIABILITY	90

After generating, deduping, and dropping the Benford variables (that were causing overfitting), I ended up with 1003 variables for feature selection.

FEATURE SELECTION:

I have tried 5 different filter & wrapper combinations that included a combination of stepwise selection (forward or backward) with a Light Gradient Boost Machine (LGBM) or Random Forest (RF).

Out of these, I ended up selecting a run with the following features

- Stepwise selection – forward
- num_filter = 250, num_wrapper = 30,
- LGBM n_estimators = 30, n_leaves = 4

Which gave me a wrapper performance of 0.72

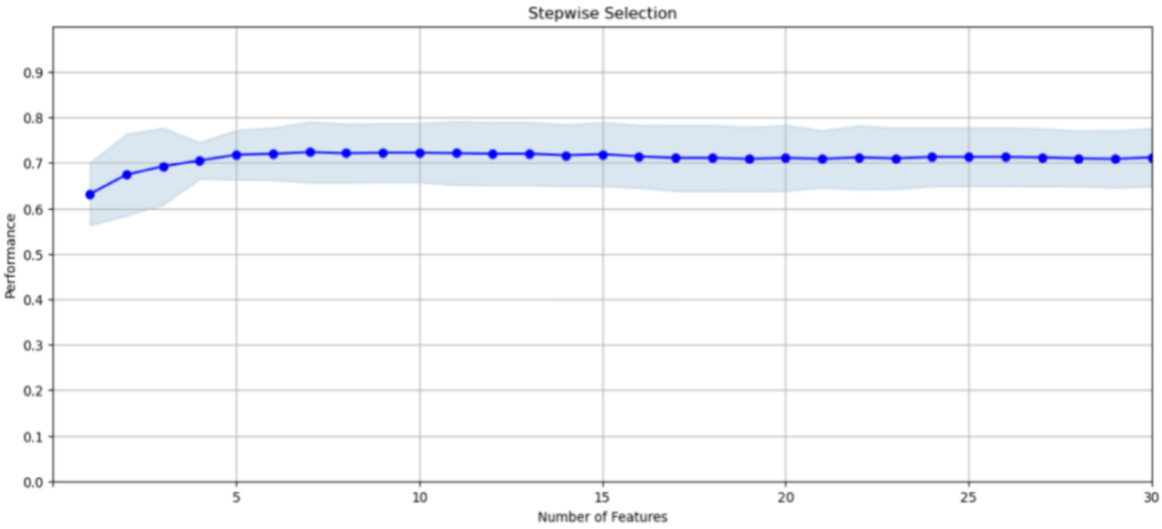


Figure 5 : Wrapper Performance Plot

The variables generated and sorted for this run are as follows:

vars_keep_sorted1

level_0	wrapper order	wrapper order	variable	0
0	1	1	Cardnum_Merchnum_MerchZip_total_14	0.6754196787184710
1	2	2	Cardnum_Merchnum_MerchZip_max_60	0.63761569548479
2	3	3	Cardnum_Merchnum_MerchZip_total_1	0.6587303118442230
3	4	4	Cardnum_Merchnum_MerchDescription_total_1	0.6103798560713620
4	5	5	Cardnum_MerchState_MerchDescription_total_3	0.6268304036875180
5	6	6	Cardnum_Merchnum_MerchDescription_total_14	0.5428754256433980
6	7	7	Cardnum_MerchZip_MerchDescription_med_1	0.561956553462122
7	8	8	Amount	0.5476782834210170
8	9	9	Cardnum_MerchState_MerchDescription_total_1	0.612652599036449
9	10	10	Merch description_total_1	0.6119349227965690
10	11	11	Cardnum_MerchState_MerchDescription_med_0	0.5380335769664980
11	12	12	Cardnum_MerchState_MerchDescription_total_14	0.5441220610101970
12	13	13	Merch description_med_0	0.5379832137215940
13	14	14	Cardnum_Merchnum_total_14	0.46033640616819800
14	15	15	Merchnum_MerchState_total_3	0.6119664143301770
15	16	16	Cardnum_Merchnum_total_3	0.59293541766767
16	17	17	Cardnum_Merchnum_total_7	0.5611380636991770
17	18	18	Merchnum_MerchState_total_14	0.47185068852880200
18	19	19	Cardnum_Merchnum_MerchZip_med_0	0.5637441295342660
19	20	20	Cardnum_MerchZip_MerchDescription_med_0	0.562566874179095
20	21	21	Cardnum_MerchState_actual/toal_14	0.4783463866779700
21	22	22	Merchnum_MerchState_med_0	0.541603115465688
22	23	23	Merch description_med_3	0.4885190948935790
23	24	24	Cardnum_Merchnum_med_0	0.5426040704526100
24	25	25	Cardnum_Merchnum_MerchZip_med_1	0.5642982412725520
25	26	26	Merch description_med_1	0.5042704898481220
26	27	27	Cardnum_MerchState_MerchDescription_med_1	0.5041445817358620
27	28	28	Merchnum_MerchState_total_7	0.5776508981164900
28	29	29	Cardnum_MerchState_actual/toal_7	0.45360555866328800
29	30	30	Cardnum_Merchnum_MerchDescription_total_0	0.5822461670698530

Figure 6 : Sorted Variables list

REASONING TO CHOOSE THE ABOVE RUN:

- The wrapper performance of the variable set from step 3 is around 0.72, better than my other trials.
- The set's variables include short-term and long-term variables with different aggregations for different combinations of my entities.
- There is no overfitting in the set.

MODEL EXPLORATION:

Models I tried varying parameters – Logistic regression, Single Decision Tree, Random Forest, Light Gradient Boost Machine, Catboost, Neural Net

MODEL	PARAMETERS							FDR at 3%		
logistic regression	**penalty**	**solver**	**C**	**max_iter**				**train**	**test**	**oot**
1	none	lbfgs (doesnt suppo	1	20				59.3	59.39	38.1
2	l2	lbfgs	1	30				59.38	59.47	38.21
3	l1	saga	1	70				59.75	59.41	37.31
4	elasticnet (l1 ratio = 0.6	saga	1	50				59.99	58.52	37.09
Single DT	**max_dept**	**min_samples_split**	**min_samples_leaf**	**max_features**				**train**	**test**	**oot**
1	10	70	50	not specified				79.46	73.61	43.07
2	50	50	100	5				78.39	73.15	46.7
3	50	60	5	7				96.87	73.73	33.01
4	50	500	100	10				73.34	70.44	45.02
5	50	300	200	not specified				71.83	66.11	42.17
6	100	600	300	10				69.42	65.92	39.1
RF	**n_estimators**	**max_depth**	**min_samples_split**	**min_samples_leaf**	**max_features**			**train**	**test**	**oot**
1	3	2	1000	500	8			64.42	63.82	49.32
2	30	50	100	50	5			83.61	78.27	54.91
3	30	70	500	100	5			76.64	73.55	55.53
4	40	30	200	70	10			79.87	77.53	55.86
5	10	30	700	100	10			74.83	72.6	56.98
LGBM	**Num_leaves**	**n_estimators**	**min_child_Samples**	**learning rate**	**max_depth**			**train**	**test**	**oot**
1	2	20	100	0.01	6			64.94	63.27	48.43
2	4	100	200	0.01	7			68.96	68.1	54.24
3	6	500	300	0.01	10			67.21	63.92	50.72
4	8	700	500	0.01	15			79.52	75.42	53.24
5	10	1000	800	0.1	10			98.6	81.05	35.97
Catboost	**depth**	**iterations**	**bootstrap_type**	**learning rate**	**l2_leaf_reg**			**train**	**test**	**oot**
1	2	70	Bernoulli	0.1	20			62.31	63	41.5
2	3	100	Bayesian	0.01	50			64.57	64.44	53.07
3	4	200	MVS	0.001	100			61.69	63.01	28.54
4	5	500	Bayesian	0.001	200			63.82	63.54	42.4
5	6	1000	Bernoulli	0.0001	500			62.39	60.01	31.39
Neural Net	**hidden_layer_sizes**	**activation**	**alpha**	**learning_rate**	**solver**	**learning_rate_**		**train**	**test**	**oot**
1	(5,)	logistic	0.1	constant	adam	0.01		63.88	62.81	44.18
2	(10,10)	logistic	0.01	adaptive	adam	0.01		69.81	70.62	51.34
3	(20,20,20)	relu	0.001	adaptive	lbfgs	0.001		77.86	73.59	43.68
4	(10,10)	relu	0.0001	constant	adam	0.001		72.01	70.89	48.93
5	(20,20,20)	logistic	0.0001	adaptive	lbfgs	0.0001		42.57	41.1	24.86

Figure 7 : Various model outputs

PERFORMANCE PLOT:

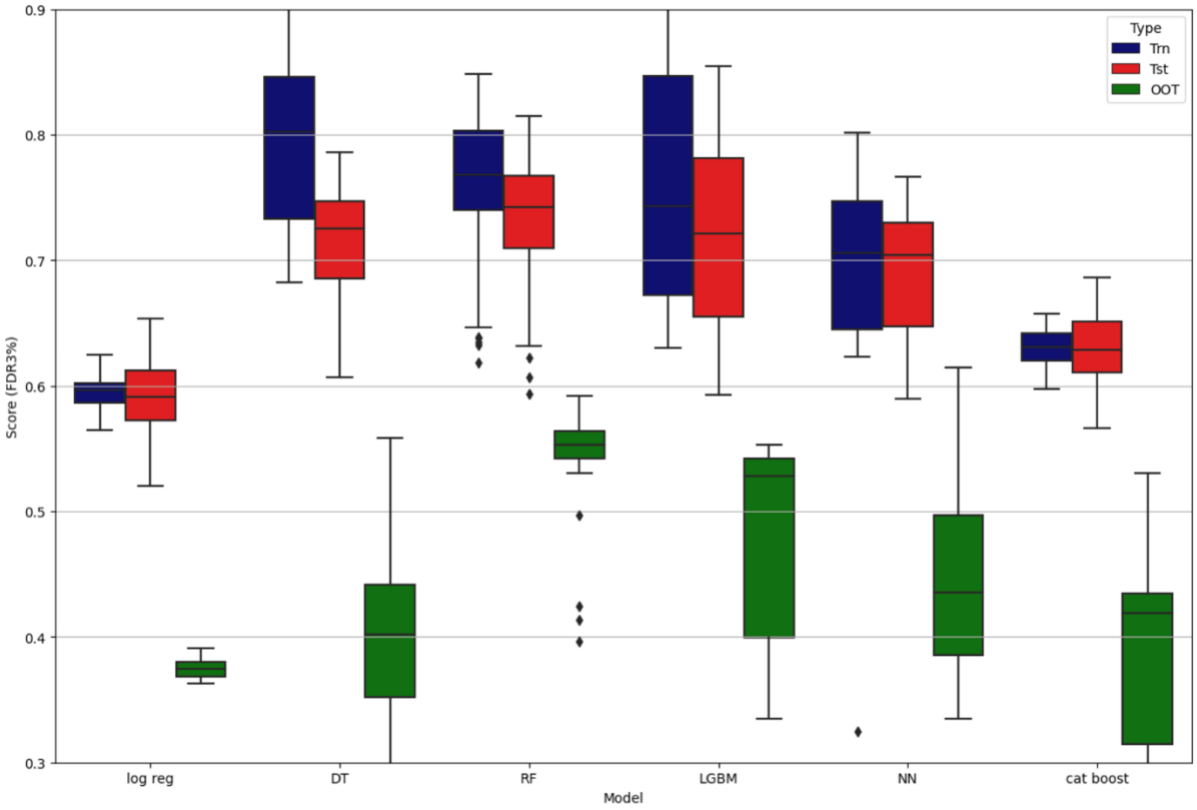


Figure 8: Model performance Comparision Plot

From the above model exploration, I ended up choosing the two best models.

1. LGBM, num_leaves = 4, n_estimators = 100, min_child_Samples = 200, learning rate = 0.01, max_Depth = 7

Train:

FDR_trn

bin	#recs	#g	#b	%g	%b	tot	cg	cb	%cg	FDR	KS	FPR
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	590.0	272.0	318.0	46.101694915254200	53.898305084745800	590.0	272.0	318.0	0.4658331906148310	51.29032258064520	50.82448939003030	0.8553459119496860
2.0	590.0	514.0	76.0	87.11864406779660	12.881355932203400	1180.0	786.0	394.0	1.3461209111149200	63.54838709677420	62.202266185659300	1.9949238578680200
3.0	590.0	538.0	52.0	91.1864406779661	8.813559322033900	1770.0	1324.0	446.0	2.267511560198660	71.93548387096770	69.66797231076910	2.968609865470850
4.0	590.0	574.0	16.0	97.28813559322030	2.711864406779670	2360.0	1898.0	462.0	3.2505566021579000	74.51612903225810	71.26557243010020	4.108225108225110
5.0	590.0	576.0	14.0	97.62711864406780	2.372881355932100	2950.0	2474.0	476.0	4.237026888165780	76.7741935483871	72.53716666022130	5.197478991596640
6.0	591.0	586.0	5.0	99.15397631133670	0.8460236886632800	3541.0	3060.0	481.0	5.240623394416850	77.58064516129030	72.34002176687350	6.361746361746360
7.0	590.0	576.0	14.0	97.62711864406780	2.372881355932100	4131.0	3636.0	495.0	6.22709368042473	79.83870967741940	73.61161599699460	7.345454545454550
8.0	590.0	573.0	17.0	97.11864406779660	2.881355932203390	4721.0	4209.0	512.0	7.208426100359650	82.58064516129030	75.37221906093070	8.220703125
9.0	590.0	583.0	7.0	98.8135593220339	1.1864406779661000	5311.0	4792.0	519.0	8.206884740537760	83.70967741935480	75.50279267881710	9.233140655105970
10.0	590.0	585.0	5.0	99.15254237288140	0.8474576271186440	5901.0	5377.0	524.0	9.208768624764520	84.51612903225810	75.30736040749360	10.261450381679400
11.0	590.0	582.0	8.0	98.64406779661020	1.355932203389830	6491.0	5959.0	532.0	10.205514642918300	85.80645161290320	75.60093696998490	11.201127819548900
12.0	590.0	588.0	2.0	99.66101694915250	0.3389830508474600	7081.0	6547.0	534.0	11.212536393218000	86.12903225806450	74.9164958648465	12.260299625468200
13.0	590.0	587.0	3.0	99.49152542372880	0.5084745762711830	7671.0	7134.0	537.0	12.217845521493400	86.61290322580650	74.39505770431310	13.28491620111730
14.0	590.0	585.0	5.0	99.15254237288140	0.8474576271186440	8261.0	7719.0	542.0	13.219729405720200	87.41935483870970	74.19962543298950	14.24169741697420
15.0	591.0	586.0	5.0	99.15397631133670	0.8460236886632800	8852.0	8305.0	547.0	14.223325911971200	88.2258064516129	74.00248053964170	15.182815356489900
16.0	590.0	589.0	1.0	99.83050847457630	0.16949152542372300	9442.0	8894.0	548.0	15.232060284295300	88.38709677419360	73.1550364898983	16.22992700729930
17.0	590.0	589.0	1.0	99.83050847457630	0.16949152542372300	10032.0	9483.0	549.0	16.240794656619300	88.54838709677420	72.3075924401549	17.273224043715800
18.0	590.0	589.0	1.0	99.83050847457630	0.16949152542372300	10622.0	10072.0	550.0	17.24952902894330	88.70967741935480	71.46014839041150	18.312727272727300
19.0	590.0	584.0	6.0	98.98305084745760	1.0169491525423700	11212.0	10656.0	556.0	18.249700291145700	89.6774193548387	71.42771906369300	19.165467625899300
20.0	590.0	586.0	4.0	99.32203389830510	0.6779661016949210	11802.0	11242.0	560.0	19.253296797396800	90.3225806451613	71.06928384776450	20.075

Figure 9 : Train performance

Test:

FDR_tst

bin	#recs	#g	#b	%g	%b	tot	cg	cb	%cg	FDR	KS	FPR
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	253.0	106.0	147.0	41.89723320158100	58.10276679841900	253.0	106.0	147.0	0.42349180982820600	56.53846153846150	56.11496972863330	0.7210884353741500
2.0	253.0	224.0	29.0	88.53754940711460	11.462450592885400	506.0	330.0	176.0	1.3184178985217700	67.6923076923077	66.37388979378590	1.875
3.0	253.0	238.0	15.0	94.07114624505930	5.928853754940720	759.0	568.0	191.0	2.2692768677586900	73.46153846153850	71.19226159377980	2.973821989528800
4.0	253.0	249.0	4.0	98.41897233201580	1.5810276679841900	1012.0	817.0	195.0	3.2640831002796600	75.0	71.73591689972030	4.18974358974359
5.0	252.0	247.0	5.0	98.01587301587300	1.9841269841269900	1264.0	1064.0	200.0	4.250898921294450	76.92307692307690	72.67217800178250	5.32
6.0	253.0	251.0	2.0	99.2094861660079	0.7905138339920940	1517.0	1315.0	202.0	5.253695565321610	77.6923076923077	72.43861212698610	6.50990099009901
7.0	253.0	250.0	3.0	98.81422924901190	1.1857707509881400	1770.0	1565.0	205.0	6.2524970035956900	78.84615384615380	72.59365684255820	7.634146341463410
8.0	253.0	246.0	7.0	97.23320158102770	2.7667984189723300	2023.0	1811.0	212.0	7.235317618857370	81.53846153846150	74.30314391960420	8.54245283018868
9.0	253.0	247.0	6.0	97.62845849802370	2.3715415019762800	2276.0	2058.0	218.0	8.222133439872150	83.84615384615380	75.62402040628170	9.440366972477060
10.0	253.0	251.0	2.0	99.2094861660079	0.7905138339920940	2529.0	2309.0	220.0	9.22493008389932	84.61538461538460	75.39045453148530	10.495454545454500
11.0	253.0	251.0	2.0	99.2094861660079	0.7905138339920940	2782.0	2560.0	222.0	10.227726727926500	85.38461538461540	75.15688865668890	11.531531531531500
12.0	253.0	253.0	0.0	100.0	0.0	3035.0	2813.0	222.0	11.238513783459800	85.38461538461540	74.14610160115550	12.67117117117120
13.0	253.0	253.0	0.0	100.0	0.0	3288.0	3066.0	222.0	12.249300838993200	85.38461538461540	73.13531454562220	13.81081081081080
14.0	253.0	253.0	0.0	100.0	0.0	3541.0	3319.0	222.0	13.260087894526600	85.38461538461540	72.12452749008880	14.95045045045050
15.0	253.0	252.0	1.0	99.60474308300400	0.3952569169960470	3794.0	3571.0	223.0	14.266879744306800	85.76923076923080	71.50235102492390	16.013452914798200
16.0	252.0	250.0	2.0	99.2063492063492	0.7936507936507980	4046.0	3821.0	225.0	15.265681182580900	86.53846153846150	71.27278035588060	16.982222222222200
17.0	253.0	251.0	2.0	99.2094861660079	0.7905138339920940	4299.0	4072.0	227.0	16.26847782660810	87.3076923076923	71.03921448108420	17.938325991189400
18.0	253.0	252.0	1.0	99.60474308300400	0.3952569169960470	4552.0	4324.0	228.0	17.275269676388300	87.6923076923077	70.41703801591940	18.964912280701800
19.0	253.0	250.0	3.0	98.81422924901190	1.1857707509881400	4805.0	4574.0	231.0	18.274071114662400	88.84615384615380	70.57208273149140	19.8008658008658
20.0	253.0	252.0	1.0	99.60474308300400	0.3952569169960470	5058.0	4826.0	232.0	19.28086296444270	89.23076923076920	69.94990626632660	20.801724137931000

Figure 10: Test Performance

Out-Of-Time:

FDR_oot

bin	#recs	#g	#b	%g	%b	tot	cg	cb	%cg	FDR	KS	FPR
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	121.0	63.0	58.0	52.06611570247930	47.93388429752070	121.0	63.0	58.0	0.5286121832522240	32.402234636871500	31.87362245361930	1.0862068965517200
2.0	121.0	103.0	18.0	85.12396694214880	14.876033057851200	242.0	166.0	76.0	1.3928511495217300	42.45810055865920	41.065249409137500	2.1842105263157900
3.0	121.0	96.0	25.0	79.33884297520660	20.661157024793400	363.0	262.0	101.0	2.1983554287632200	56.42458100558660	54.226225576823400	2.594059405940590
4.0	121.0	120.0	1.0	99.17355371900830	0.8264462809917320	484.0	382.0	102.0	3.2052357778150700	56.98324022346370	53.77800444564860	3.7450980392156900
5.0	121.0	121.0	0.0	100.0	0.0	605.0	503.0	102.0	4.220506796442360	56.98324022346370	52.76273342702130	4.931372549019610
6.0	121.0	118.0	3.0	97.52066115702480	2.47933884297521	726.0	621.0	105.0	5.210605806343350	58.659217877095000	53.44861207075160	5.914285714285720
7.0	121.0	113.0	8.0	93.38842975206610	6.611570247933880	847.0	734.0	113.0	6.158751468367180	63.12849162011170	56.96974015174460	6.495575221238940
8.0	121.0	121.0	0.0	100.0	0.0	968.0	855.0	113.0	7.174022486994460	63.12849162011170	55.95446913311730	7.566371681415930
9.0	121.0	118.0	3.0	97.52066115702480	2.47933884297521	1089.0	973.0	116.0	8.164121496895450	64.80446927374300	56.640347776847600	8.387931034482760
10.0	121.0	118.0	3.0	97.52066115702480	2.47933884297521	1210.0	1091.0	119.0	9.154220506796440	66.4804469273743	57.32622642057790	9.168067226890760
11.0	121.0	118.0	3.0	97.52066115702480	2.47933884297521	1331.0	1209.0	122.0	10.144319516697400	68.15642458100560	58.012105064308100	9.90983606557377
12.0	121.0	117.0	4.0	96.69421487603310	3.305785123966940	1452.0	1326.0	126.0	11.12602785702300	70.39106145251400	59.265033595491000	10.523809523809500
13.0	121.0	120.0	1.0	99.17355371900830	0.8264462809917320	1573.0	1446.0	127.0	12.132908206074800	70.94972067039110	58.816812464316200	11.385826771653500
14.0	121.0	120.0	1.0	99.17355371900830	0.8264462809917320	1694.0	1566.0	128.0	13.1397885551267	71.50837988826820	58.36859133314150	12.234375
15.0	121.0	118.0	3.0	97.52066115702480	2.47933884297521	1815.0	1684.0	131.0	14.129887565027700	73.18435754189940	59.05446997687180	12.854961832061100
16.0	121.0	119.0	2.0	98.34710743801650	1.652892561983480	1936.0	1803.0	133.0	15.128377244504100	74.30167597765360	59.17329873314950	13.556390977443600
17.0	120.0	120.0	0.0	100.0	0.0	2056.0	1923.0	133.0	16.135257593556000	74.30167597765360	58.16641838409770	14.458646616541400
18.0	121.0	121.0	0.0	100.0	0.0	2177.0	2044.0	133.0	17.15052861218330	74.30167597765360	57.15114736547040	15.368421052631600
19.0	121.0	116.0	5.0	95.86776859504130	4.132231404958670	2298.0	2160.0	138.0	18.12384628293340	77.09497206703910	58.97112578410570	15.652173913043500
20.0	121.0	118.0	3.0	97.52066115702480	2.47933884297521	2419.0	2278.0	141.0	19.11394529283440	78.77094972067040	59.657004427836000	16.156028368794300

Figure 11: OOT Performance

FINANCIAL PLOT:

MAX SAVINGS:

Using the LGBM model with the above-mentioned parameters, gave me the following financial plot and max savings of $21,288,000.

```
Max possible savings: 21,288,000.0
```

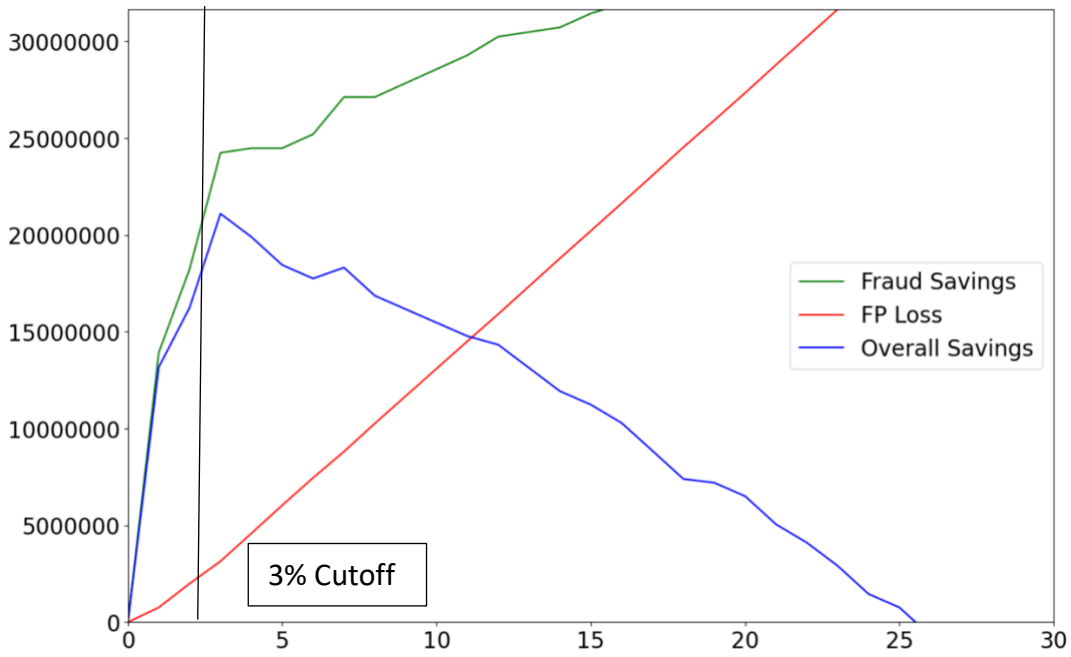


Figure 12 : Financial Plot

SUMMARY:

The credit card transaction data is analyzed following a series of steps such as data cleaning, feature engineering, feature selection, and model building, out of which the best model is selected to generate a financial plot with an FDR (Fraud Detection rate) cut off at 3%.

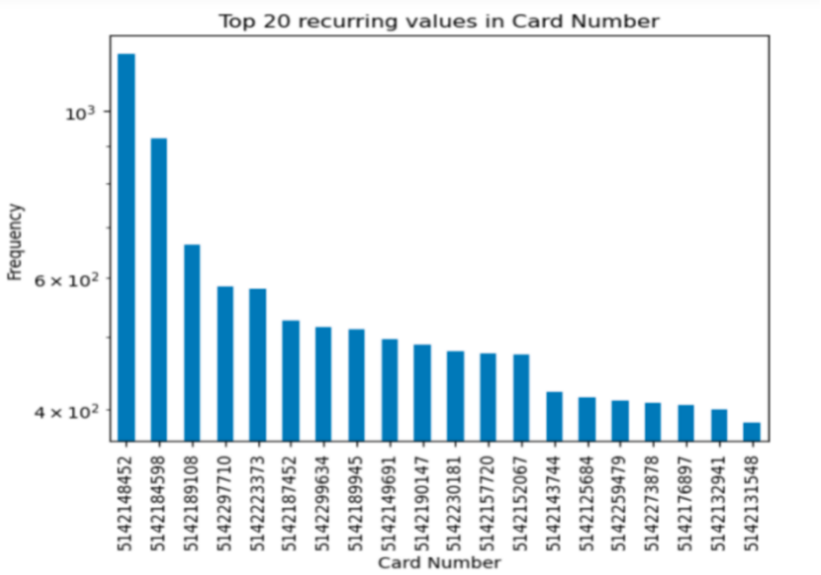
I wanted to experiment with two models, which I selected based on the criteria of the trn and tst values not being very different from each other and the oot being considerably higher than other models.

Based on this, I selected LGBM and NN models. But after generating the financial plot for both and observing the max saving for each of them, I conclude that the LGBM did a better job.

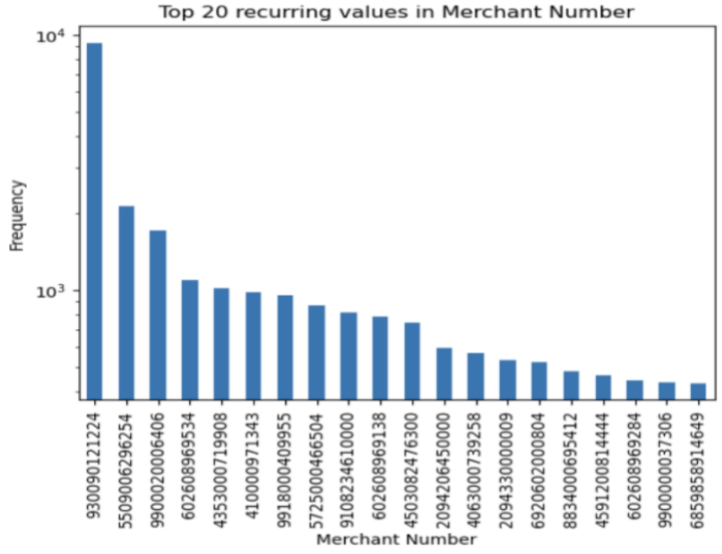
APPENDIX:

VISUALIZATIONS:

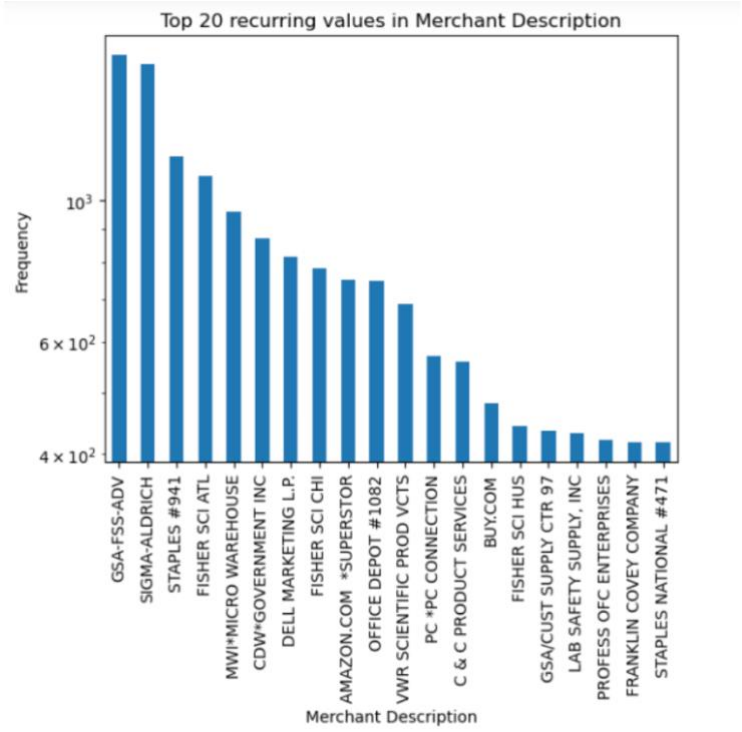
a) *Record number* – values from 1 to 96753, positive and unique
b) *Card Number* – The distribution indicates the top 20 field values of Cardnum, where the number with max count of 1,192 is 5142148452.



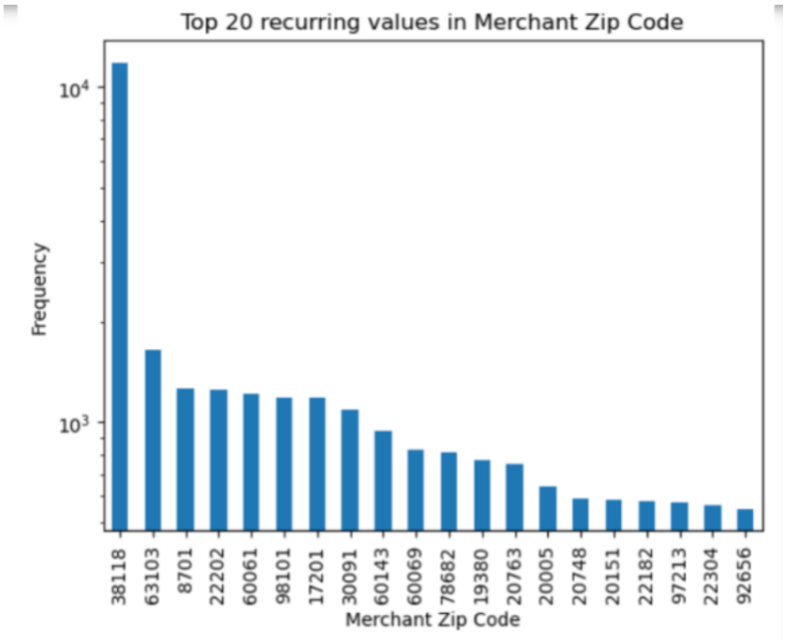
c) *Merchant Number* – The distribution indicates the top 20 field values of the merchnum, with the most recurring merchnum of 930090121224 with a count of 9310.



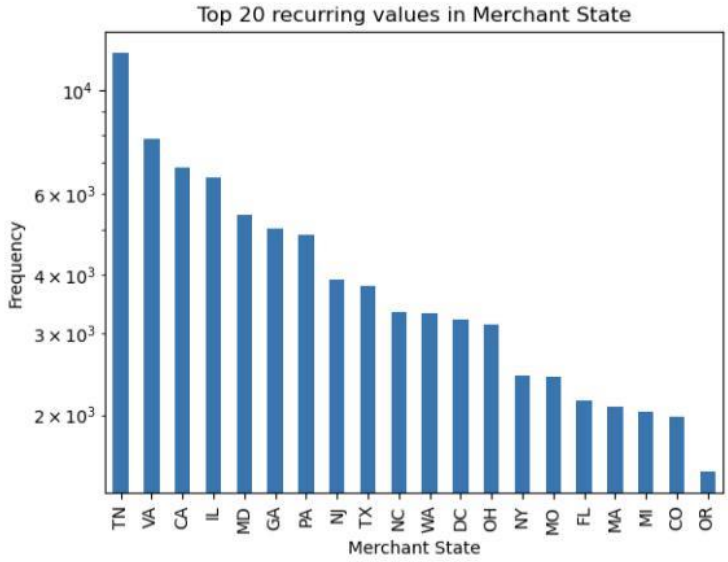
d) *Merchant Description* – The distribution indicates the top 20 field values of the merch description field. Mostly used is GSA-FSS-ADV with the count of 1688.



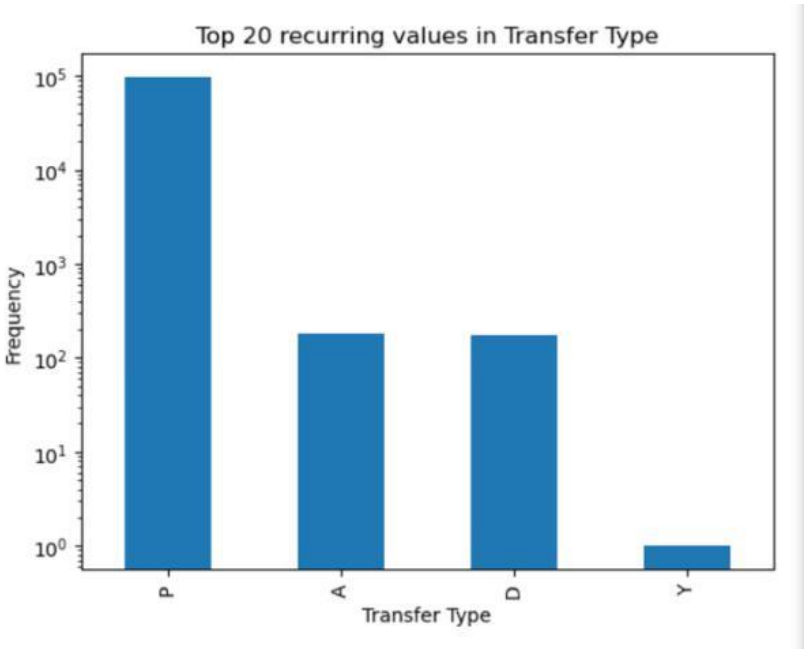
f) Merchant zip– represents the zip code of the merchant. Most merchants are from the zip code of 38118, with the highest count of 11,868



e) *Merchant state*– Represents the state of merchant in merch state. Max merchants are from state of Tennessee with phone number count of 12,035.

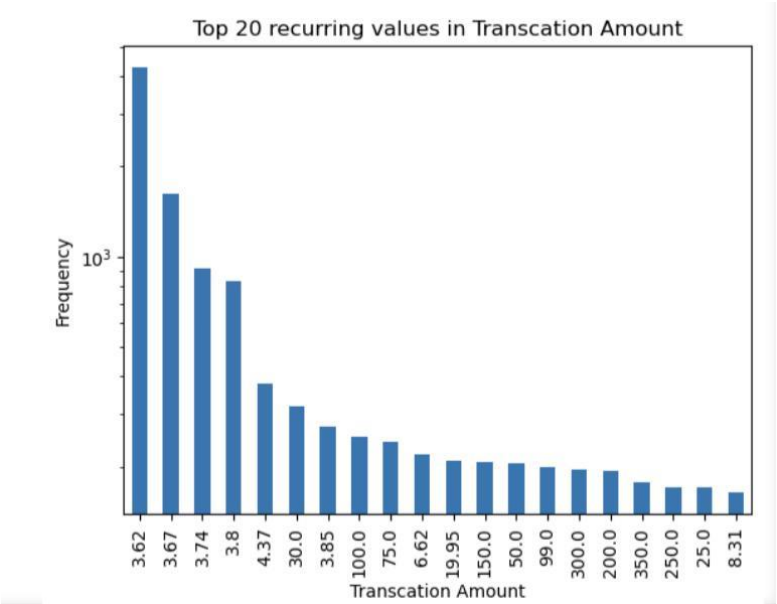


g) Transtype– Represents four different Transaction types. The most used is P, with count of 96,398.



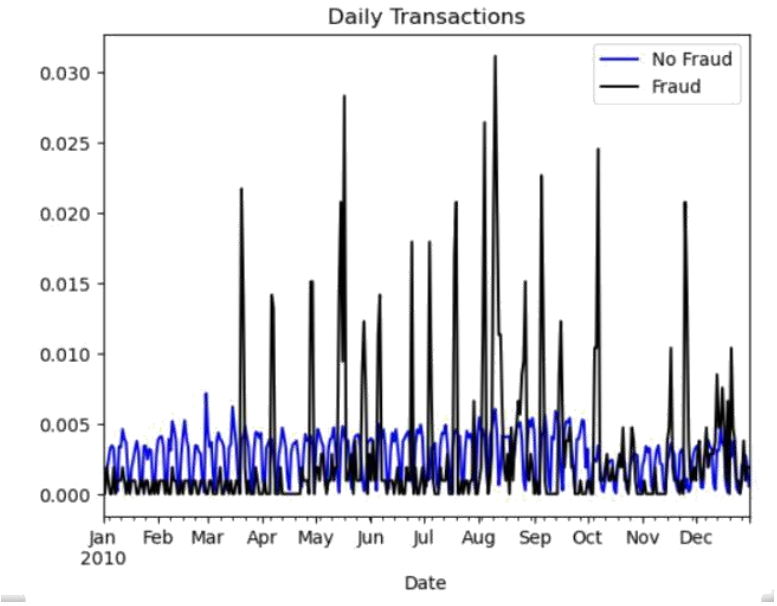
h) Fraud– indicates the fraud levels (0-No, 1-Yes) with count of 0 – 95,694 and 1 – 1,059

i) Amount – The transaction amount, with most transactions of 3.62, with the count of 4,283

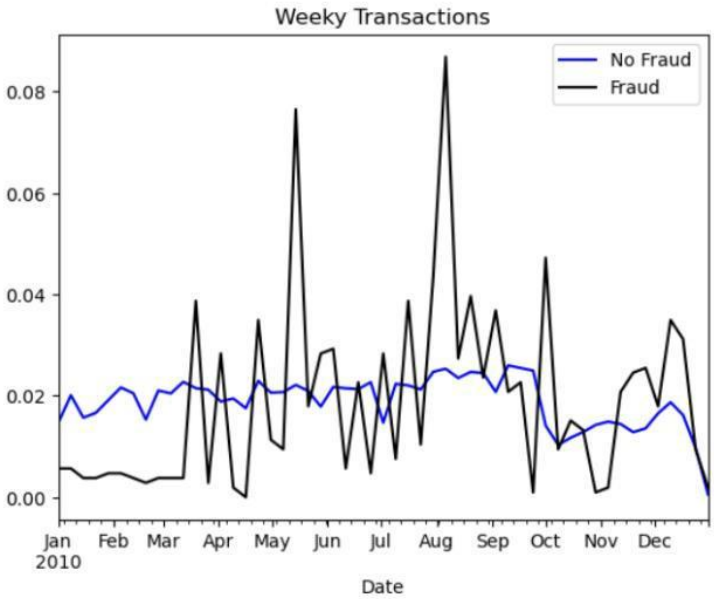


j) Based on Fraud Label:

(1) Daily Transactions: Most fraud transactions occurred in August on a daily basis.

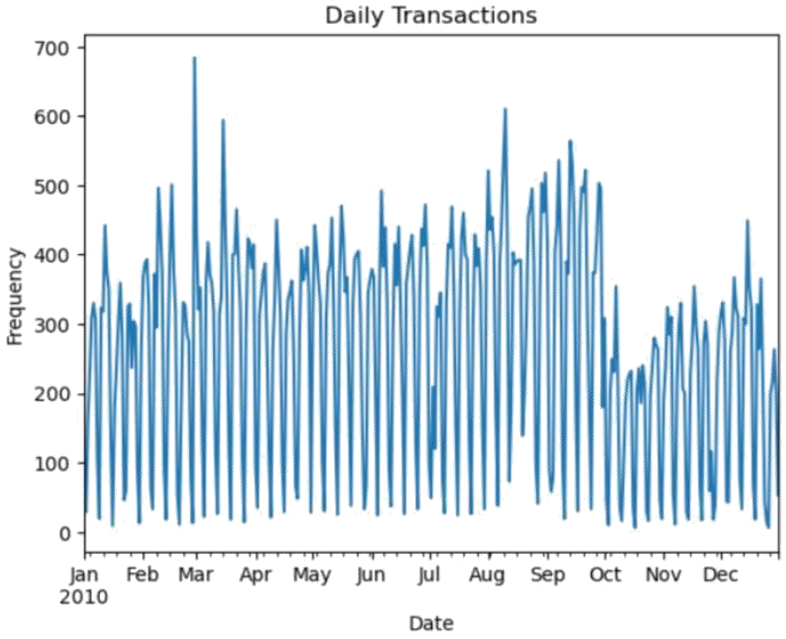


(2) Weekly Transactions: Most fraud transactions occurred in August per given

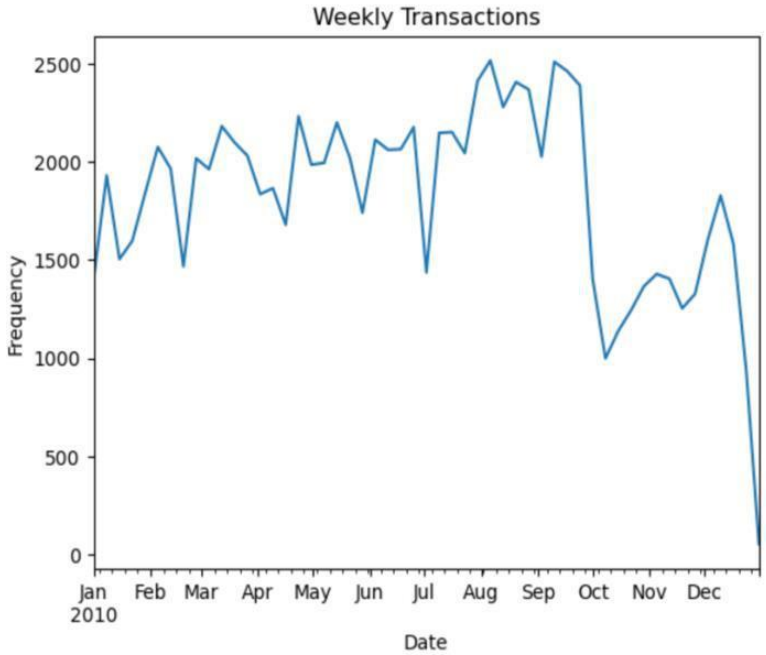


k) Based on Date:

(1) Daily Transactions – Date(Day) vs Frequency of Transactions : Most number of transactions are on 28th February 2010, with the count of 684.



2) Weekly Transactions: Date(Week) vs Frequency of Transactions : Most number of transactions are in the first week of Aug, with count of 2,516.



(3) Monthly Transactions: Date(Month) vs Frequency of Transactions : Most number of transactions are in the month of August, with a count of 10,199.