FRAUD DETECTION WITH MACHINE LEARNING TECHNIQUES

Caner Dağdaş - 150716001
Ceyhun Vardar - 150317022
Büşra Gökmen - 150116027
Cem Güleç - 150117828
Ömer Faruk Çakı -150117821

CSE4062 - Group 8

Goal:

Our project will take the financial sector one step further by identifying fraud, which is the bleeding wound of the financial sector, through machine learning by modeling in a way to detect fraud using the data set we have obtained.

Story of Dataset

- Fraud detection research in the financial domain suffers from a serious problem which is the lack of public available data for the testing, evaluation and comparison of results using standard data sets
- Our dataset is generated dataset to provide source about this topic.
- Original logs are provided from multinational company which operates over 10 countries.

Topics

- Data description
- Data Analysis
- K-Mean Clustering
- Classification Methods
- Conclusion

Data Description

Feature Names	Description	Type	Missing Value (%)
step	maps a unit of time in the real world	numeric	0
type	CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER	nominal	0
amount	amount of the transaction in local currency	numeric	0
nameOrig	customer who started the transaction	nominal	0
oldbalanceOrg	initial balance before the transaction	numeric	0
newbalanceOrig	new balance after the transaction	numeric	0
nameDest	customer who is the recipient of the transaction	nominal	0
oldbalanceDest	initial balance recipient before the transaction	numeric	0
newbalanceDest	new balance recipient after the transaction	numeric	0
isFraud	This is the transactions made by the fraudulent agents inside the simulation	nominal	0
isFlaggedFraud	The business model aims to control massive transfers from one account to another and flags illegal attempts	nominal	0

Table 1 - Features of dataset and missing values

Data Analysis

Feature Names	Description	Average	Standard Deviation	Entropy	# of Values
step	maps a unit of time in the real world	243.397	142.332	5.5276	6.362.622
type	CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER	-	-	1.3077	6.362.622
amount	amount of the transaction in local currency	179861.90	603858.2	15.4085	6.362.622
nameOrig	customer who started the transaction	-	-	15.6639	6.362.622
oldbalanceOrg	initial balance before the transaction	833883.10	2888243	9.3923	6.362.622
newbalanceOrig	new balance after the transaction	855113.66	2924049	7.0845	6.362.622
nameDest	customer who is the recipient of the transaction	-	-	14.0318	6.362.622
oldbalanceDest	initial balance recipient before	1100701.66	3399180	9.3598	6.362.622
newbalanceDest	new balance recipient after the transaction	1224996.39	3674129	9.9373	6.362.622
isFraud	This is the transactions made by the fraudulent agents inside the simulation	-	-	0.0098	6.362.622
isFlaggedFraud	The business model aims to control massive transfers from one account to another and flags illegal attempts	-	-	0.000034	6.362.622

Table 2 - Features of dataset and descriptive values

Data Analysis

Feature Names	Description	Type	Min. or Least Frequent	Max. or Most Frequent
step	maps a unit of time in the real world	numeric	1	743
type	CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER	nominal	DEBIT	CASH_OUT
amount	amount of the transaction in local currency	numeric	0	92.445.516,64
nameOrig	customer who started the transaction	nominal	C2066766136	C1999539787
oldbalanceOrg	initial balance before the transaction	numeric	0	59.585.040,37
newbalanceOrig	new balance after the transaction	numeric	0	49.585.040,37
nameDest	customer who is the recipient of the transaction	nominal	M917557255	C1286084959
oldbalanceDest	initial balance recipient before	numeric	0	356.015.889,35
newbalanceDest	new balance recipient after the transaction	numeric	0	356.179.278,92
isFraud	This is the transactions made by the fraudulent agents inside the simulation	nominal	0	1
isFlaggedFraud	The business model aims to control massive transfers from one account to another and flags illegal attempts	nominal	0	1

Table 3 - Features of dataset and descriptive values

Data Analysis

- We find that of the five types of transactions. Fraudness occured only in 2 types named as TRANSFER and CASH_OUT. That’s why we reduced our instances to 2.7 million.

Dataset Sample

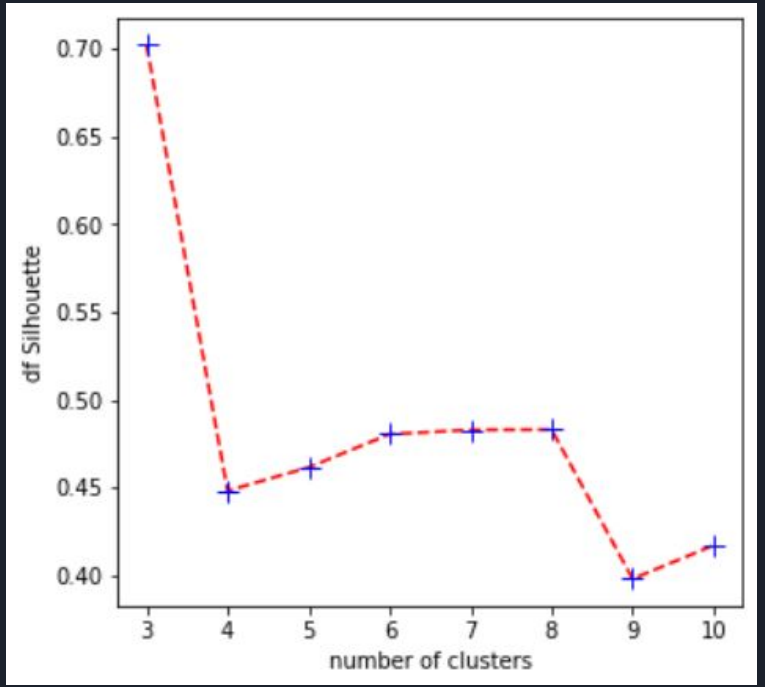
step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0
1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0	0
1	TRANSFER	181.0	C1305486145	181.0	0.0	C553264065	0.0	0.0	1	0
1	CASH_OUT	181.0	C840083671	181.0	0.0	C38997010	21182.0	0.0	1	0
1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0	0
1	PAYMENT	7817.71	C90045638	53860.0	46042.29	M573487274	0.0	0.0	0	0
1	PAYMENT	7107.77	C154988899	183195.0	176087.23	M408069119	0.0	0.0	0	0
1	PAYMENT	7861.64	C1912850431	176087.23	168225.59	M633326333	0.0	0.0	0	0
1	PAYMENT	4024.36	C1265012928	2671.0	0.0	M1176932104	0.0	0.0	0	0
1	DEBIT	5337.77	C712410124	41720.0	36382.23	C195600860	41898.0	40348.79	0	0
1	DEBIT	9644.94	C1900366749	4465.0	0.0	C997608398	10845.0	157982.12	0	0
1	PAYMENT	3099.97	C249177573	20771.0	17671.03	M2096539129	0.0	0.0	0	0
1	PAYMENT	2560.74	C1648232591	5070.0	2509.26	M972865270	0.0	0.0	0	0
1	PAYMENT	11633.76	C1716932897	10127.0	0.0	M801569151	0.0	0.0	0	0
1	PAYMENT	4098.78	C1026483832	503264.0	499165.22	M1635378213	0.0	0.0	0	0
1	CASH_OUT	229133.94	C905080434	15325.0	0.0	C476402209	5083.0	51513.44	0	0
1	PAYMENT	1563.82	C761750706	450.0	0.0	M1731217984	0.0	0.0	0	0
1	PAYMENT	1157.86	C1237762639	21156.0	19998.14	M1877062907	0.0	0.0	0	0
1	PAYMENT	671.64	C2033524545	15123.0	14451.36	M473053293	0.0	0.0	0	0
1	TRANSFER	215310.3	C1670993182	705.0	0.0	C1100439041	22425.0	0.0	0	0
1	PAYMENT	1373.43	C20804602	13854.0	12480.57	M1344519051	0.0	0.0	0	0
1	DEBIT	9302.79	C1566511282	11299.0	1996.21	C1973538135	29832.0	16896.7	0	0
1	DEBIT	1065.41	C1959239586	1817.0	751.59	C515132998	10330.0	0.0	0	0
1	PAYMENT	3876.41	C504336483	67852.0	63975.59	M1404932042	0.0	0.0	0	0
1	TRANSFER	311685.89	C1984094095	10835.0	0.0	C932583850	6267.0	2719172.89	0	0
1	PAYMENT	6061.13	C1043358826	443.0	0.0	M1558079303	0.0	0.0	0	0
1	PAYMENT	9478.39	C1671590089	116494.0	107015.61	M58488213	0.0	0.0	0	0
1	PAYMENT	8009.09	C1053967012	10968.0	2958.91	M295304806	0.0	0.0	0	0
1	PAYMENT	8901.99	C1632497828	2958.91	0.0	M33419717	0.0	0.0	0	0
1	PAYMENT	9920.52	C764826684	0.0	0.0	M1940055334	0.0	0.0	0	0
1	PAYMENT	3448.92	C2103763750	0.0	0.0	M335107734	0.0	0.0	0	0

Descriptive Analysis-K-means Clustering

Silhouette: A measure of how similar an object is to its own cluster compared to other clusters.

Different number of clusters are experimented.
To find best possible silhouette score.

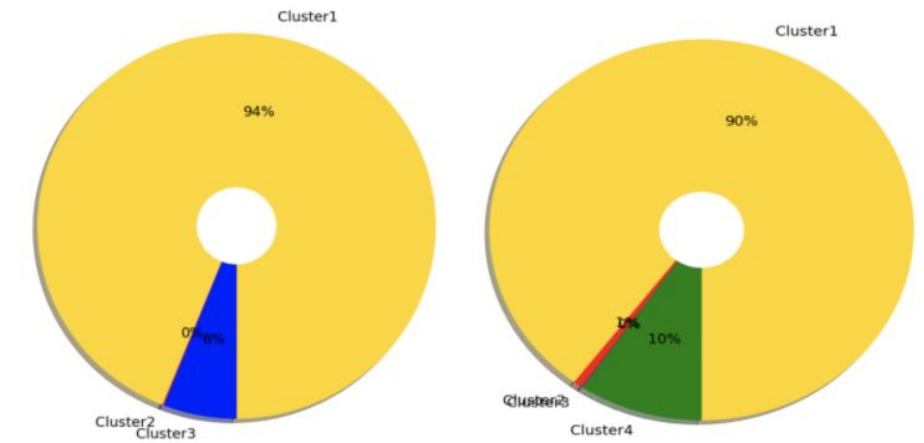
In our trials, we obtained the highest silhouette score (0.70) at k=3. In other words, the experiment in which the data is most consistently distributed to the clusters is the experiment where the k value is 3. The lowest score is the experiment with 0.39 and k=9.



Descriptive Analysis-K-means Clustering

#	feature name	description	type	overall avg./ mode	cluster 1	cluster 2	cluster 3	cluster 4
	step	maps a unit of time in the real world	numeric	243.397	1117466	463432	10474462	142065
	amount	amount of the transaction in local currency	numeric	179861.9	2648758	9141	111705	805
	nameOrig	customer who started the transaction	nominal	C1999539787	692293	690546	699864	687706
	oldbalanceOrg	initial balance before the transaction	numeric	833883.1	2657750	985	69	111605
	newbalanceOrig	new balance after the transaction	numeric	855113.6	2707346	842	59	62162
	nameDest	customer who is the recipient of the transaction	nominal	C1286084959	689529	690622	697683	692575
	oldbalanceDest	initial balance recipient before the transaction	numeric	1100701.66	2488963	19153	1099	261194
	newbalanceDest	new balance recipient after the transaction	numeric	1224996.39	2481277	19568	1350	268214

Descriptive Analysis-K-means Clustering



Graph 1 and 2: Pie chart of instance distribution when k=3 and k=4

Experiment	# of clusters	avg. number of instances in clusters	std.dev.	SSE	NMI	Silhouette Value	RI
1	k=3	0:2602389 1:2863 2:165157	0.47	4.7	0.00035	0.70	-0.0032
2	k=4	0: 2486468 1: 19187 2: 1182 3: 263572	0.88	3.39	0.000177	0.44	-0.00209
3	k=5	0: 1252392 1: 1180 2: 246299 3: 18738 4: 1251800	1.90	2.76	6.61e-05	0.46	-0.000292
4	k=6	0: 1158884 1: 1039 2: 11616 3: 352473 4: 86594 5: 1159803	2.30	2.21	0.00011	0.48	-0.000414

Predictive Analysis-Classification

- At this stage, we need to guess whether the money transfer made on a certain date is an example of fraud or not.

Predictive Analysis-Classification

Steps of Classification

- Train
- Test
- Evaluate

Predictive Analysis-Classification

ML algorithms that We have Used

- **KNN**
  KNN works on a principle assuming every data point falling in near to each other is falling in the same class. In other words it classifies a new data point based on similarity.

- **Naive Bayes Classifier**
  Bayes Optimal Classifier is a probabilistic model that finds the most probable prediction using the training data and space of hypotheses to make a prediction for a new data instance.

- **Decision Tree Classifier**
  Grow the tree: splitting the space by setting rules
  Prune the tree: removing the unnecessary splits
  Assign the class: using the class with majority votes as the prediction

- **MLP**
  MLPClassifier stands for Multi-layer Perceptron classifier which in the name itself connects to a Neural Network. MLPClassifier relies on an underlying Neural Network to perform the task of classification.

Predictive Analysis-Classification

Cross Validation

We used 3 techniques for cross validation:

1. Train-Test Split (changing parameter test_size)

   Test size: 10% of data

2. K-Fold (changing parameter k)

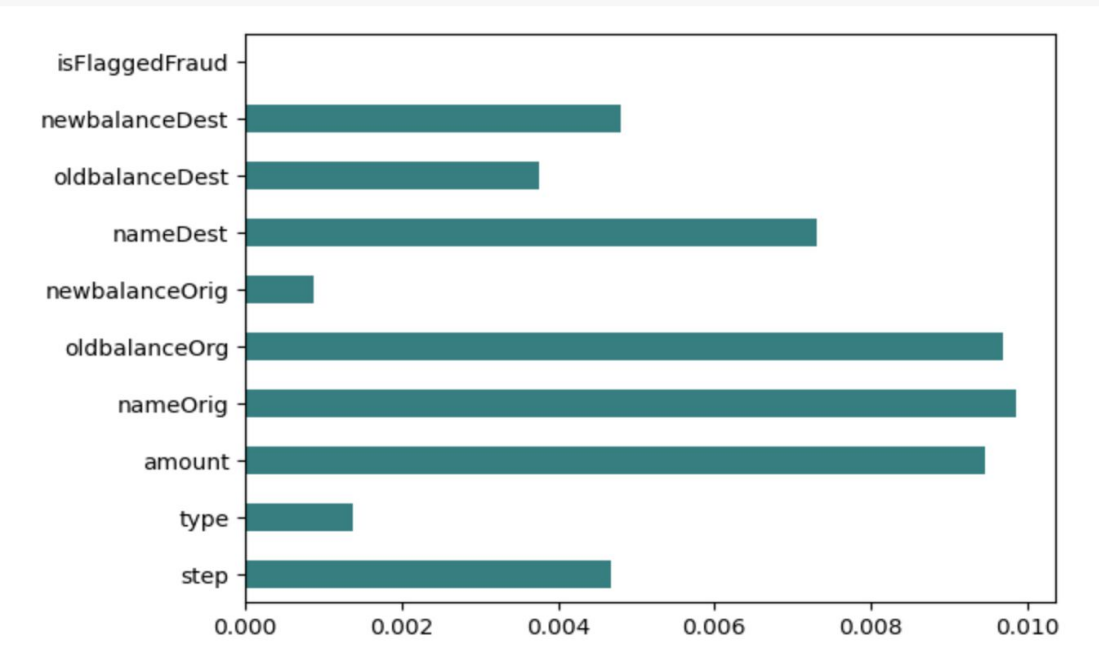
   We used 10 fold for cross validation

3. Shuffle (shuffle dataset for every fold)

Predictive Analysis-Classification

Feature Selection with Information Gain

Most important feature is nameOrig and less important feature is newbalanceOrig by this method. So we removed newbalanceOrig from our dataset.

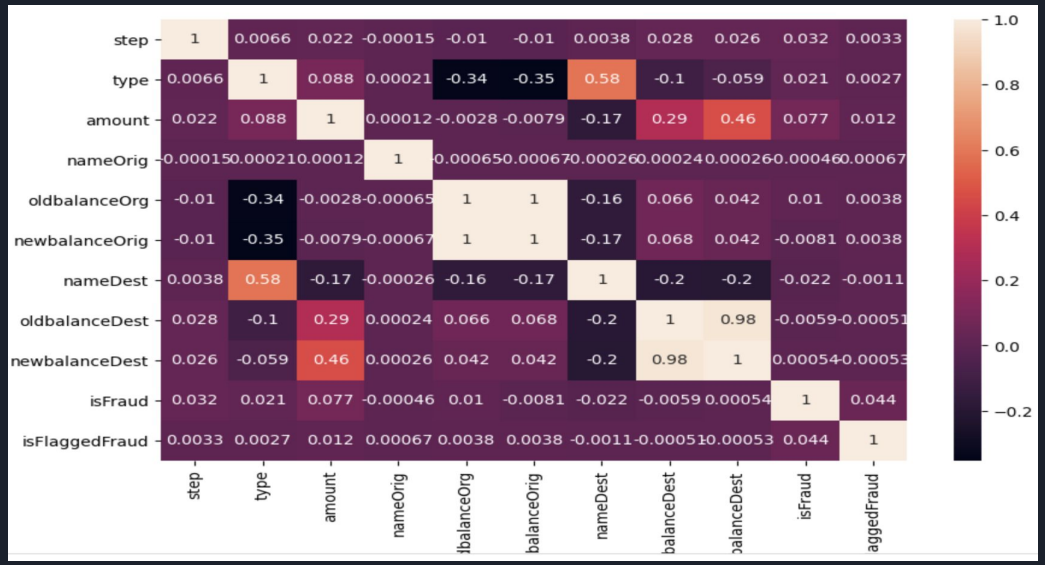


Predictive Analysis-Classification

Feature Selection with Correlation Coefficient

The logic behind using correlation for feature selection is that the good variables are highly correlated with the target. Furthermore, variables should be correlated with the target but should be uncorrelated among themselves.

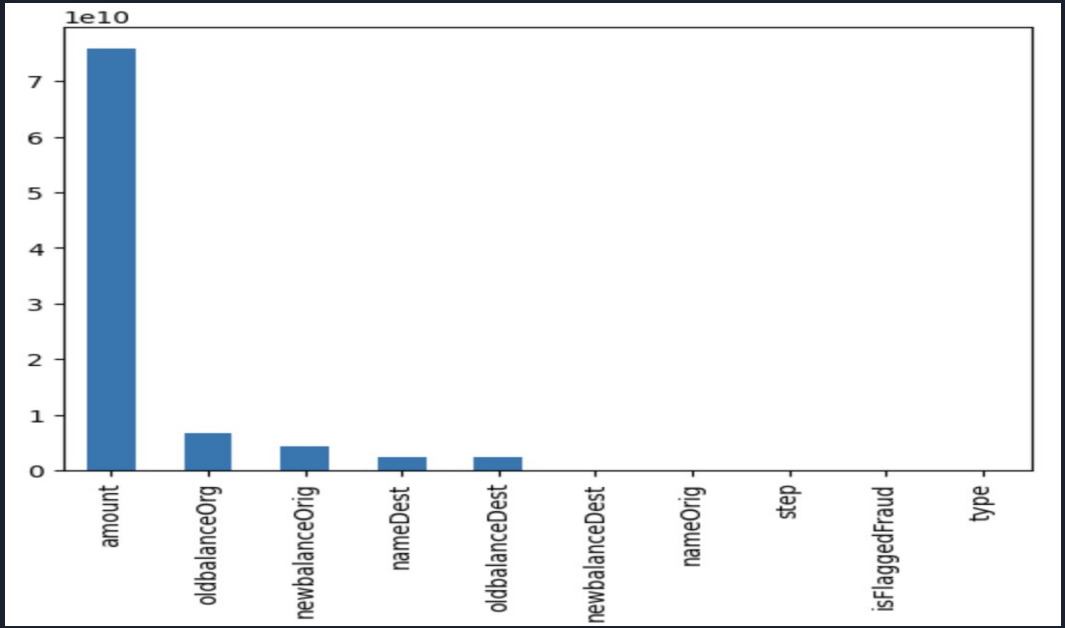
Less important features are oldbalanceDest, newbalanceOrig and nameDest. So we removed these features from our dataset.



Predictive Analysis-Classification

Feature Selection with Chi Square Test Feature

Most important feature is amount and less important feature are type, step, nameorig, is FlaggedFraud and newbalanceDest by this method. So we removed these feature from our dataset.



Predictive Analysis-Classification

Results Of Experiments

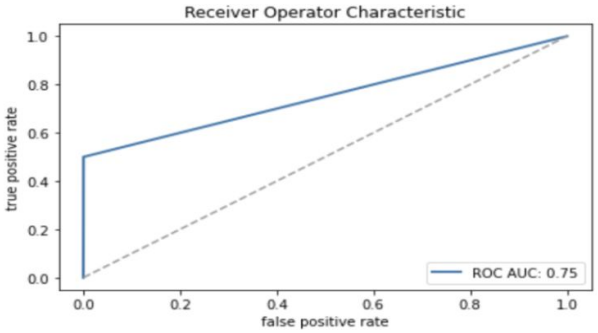
We have 18 different model experiments in this project. We evaluate our models by F1-Score, Precision, Recall, Accuracy and AUC score.

#	Experiments	F1-Score	Precision Score	Recall Score	Accuracy Score	AUC Score
		Avg	Avg	Avg	Avg	Avg
1	KNN5 IG	0.8154848101738795	0.9409437379980392	0.7498914223669924	0.9983399494767231	0.7498914223669924
2	KNN5 CC	0.840885562831088	0.9312901166551205	0.7873552298226565	0.9984843016961386	0.7873552298226565
3	KNN5 CT	0.8215544496522468	0.9409617754689756	0.7561414223669924	0.998376037531577	0.7561414223669924
4	KNN30 IG	0.7710929375909077	0.9641328694779251	0.6999457111834962	0.998159509202454	0.6999457111834962
5	KNN30 CC	0.7898426224385829	0.9233715881332109	0.7248914223669924	0.9981955972573078	0.7248914223669924
6	KNN30 CT	0.7768712017749426	0.9666509134581999	0.7061957111834962	0.9981955972573078	0.7061957111834962
7	NB IG	0.5794363740446423	0.5505534666310475	0.7044239780979186	0.9874285755537989	0.7044239780979186
8	NB CC	0.5624448718361693	0.5388830197375291	0.6927692310168482	0.9845297266469584	0.6927692310168482
9	NB CT	0.5624611556892883	0.538892527825947	0.6928297703264119	0.984529365689555	0.6928297703264119
10	DTREE ENTROPY IG	0.8912501843652416	0.9868855956461922	0.8272586391401371	0.9989261517248349	0.8472586391401371
11	DTREE ENTROPY CC	0.8912213289903483	0.9868017626103001	0.8272584581250042	0.9989257907674315	0.8472584581250042
12	DTREE ENTROPY CT	0.8912213289903483	0.9868017626103001	0.8272584581250042	0.9989257907674315	0.8472584581250042
13	DTREE GINI IG	0.8950951903423794	0.9893396080954238	0.8315264393216549	0.9989608036355629	0.8315264393216549
14	DTREE GINI CC	0.8950951903423794	0.9893396080954238	0.8315264393216549	0.9989608036355629	0.8315264393216549
15	DTREE GINI CT	0.8950951903423794	0.9893396080954238	0.8315264393216549	0.9989608036355629	0.8315264393216549
16	MLP IG	0.7879411340823693	0.9913596052077562	0.7039245085201133	0.9982060417050184	0.7039245085201135
17	MLP CC	0.8264447652988596	0.987274904015663	0.7456663619457062	0.9984394608258946	0.7456663619457062
18	MLP CT	0.8240480349705437	0.9926672311279199	0.7416603575767171	0.9984322416778263	0.7416603575767172

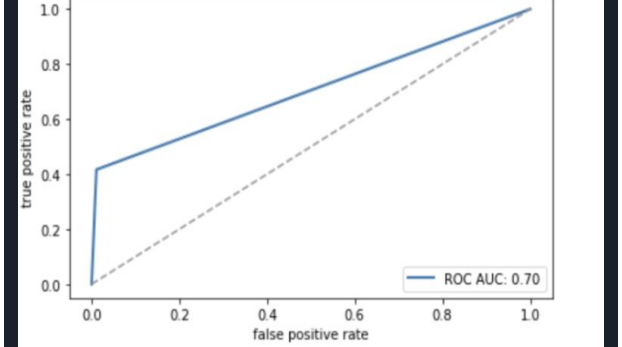
Predictive Analysis-Classification

ROC Curves

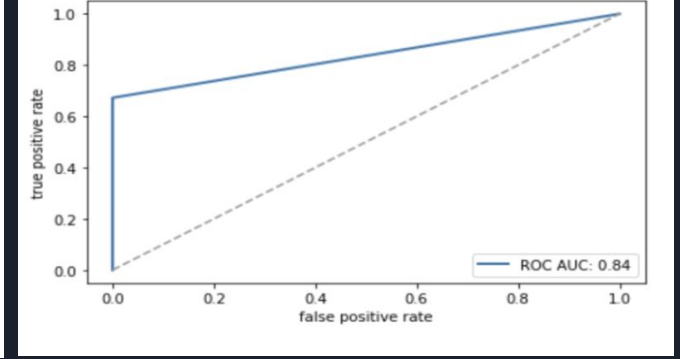
KNN5-IG



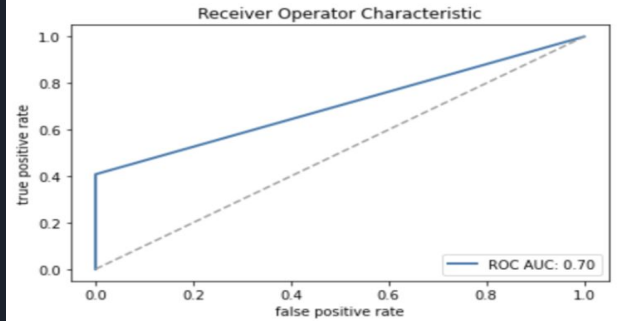
NG-IG



DTREE-ENTROPY-IG



MLP-IG



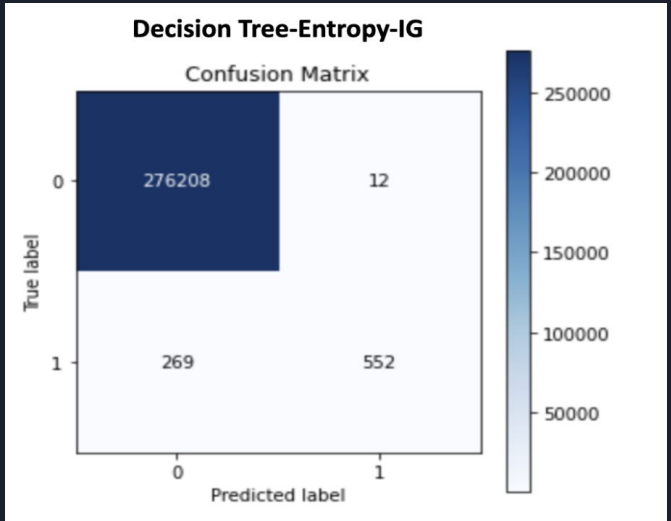
Predictive Analysis-Classification

Best Model

Our aim is to maximize these scores. The model has the maximum scores calculated above can be considered as the best model.

- After experiments, best model is chosen, according to 6 different metrics, as decision tree entropy with information gain.

Confusion matrix of the best fold performance of the our best performing model:



Predictive Analysis-Classification

T-Test Between Two Best Model

For t-test, we chose our top two experiments according to F1-Score and ROC-AUC metric :

1. Decision Tree-Entropy-IG

- F1-Score: 0.89
- ROC-AUC metric: 0.84

2. Decision Tree-Gini-IG

- F1-Score: 0.89
- ROC-AUC metric: 0.83

We compared them by using 5 different metrics(Accuracy, P, R, F1, AUC) according to T-Test. According to all score metric types there's a significant difference between models. So we can concluded that, Decision Tree-Entropy-IG is our best experiment.

For Example:

- **F1 T_Test** between Decision Tree-Gini-IG & Decision Tree-Entropy-IG:

  T Value = [7.561632004604 1725], P Value = [3.9745984281580604e-12]

Conclusion

- According to the elbow plot, the optimum k value is 4 according to the k-means clustering model.
- According to all score matric types there's a significance difference between models. we can concluded that Decision Tree-Entropy-IG is our best experiment.
- Our decision tree classifier model performed better than our deep learning model.
- In general, the Correlation coefficient and Chi-square test feature selection methods resulted in higher accuracy than the Information gain method in classifier models.

Thank you for listening!