EXCELR
Raising Excellence

Incident Impact Prediction

Presented by Group - 4

Mrs.Bhanupriya

01/04/2021

Business Problem:

To predict the impact of the incident raised by the customer.

Objective:

The Objective of resolution is to predict the impact of the incident built by the customer and to develop the accuracy of model as well to figure out the important features combined with it.

Dataset Used:

Incident_event_log_dataset

Project Architecture / Project Flow

EXCELR
Raising Excellence

Research about Incident Impact Prediction

Exploratory Data Analysis (EDA)

Data Cleaning

Feature Selection & Data Balancing using SMOTE

Model Building

Deployment

Data set details

EXCELR
Raising Excellence

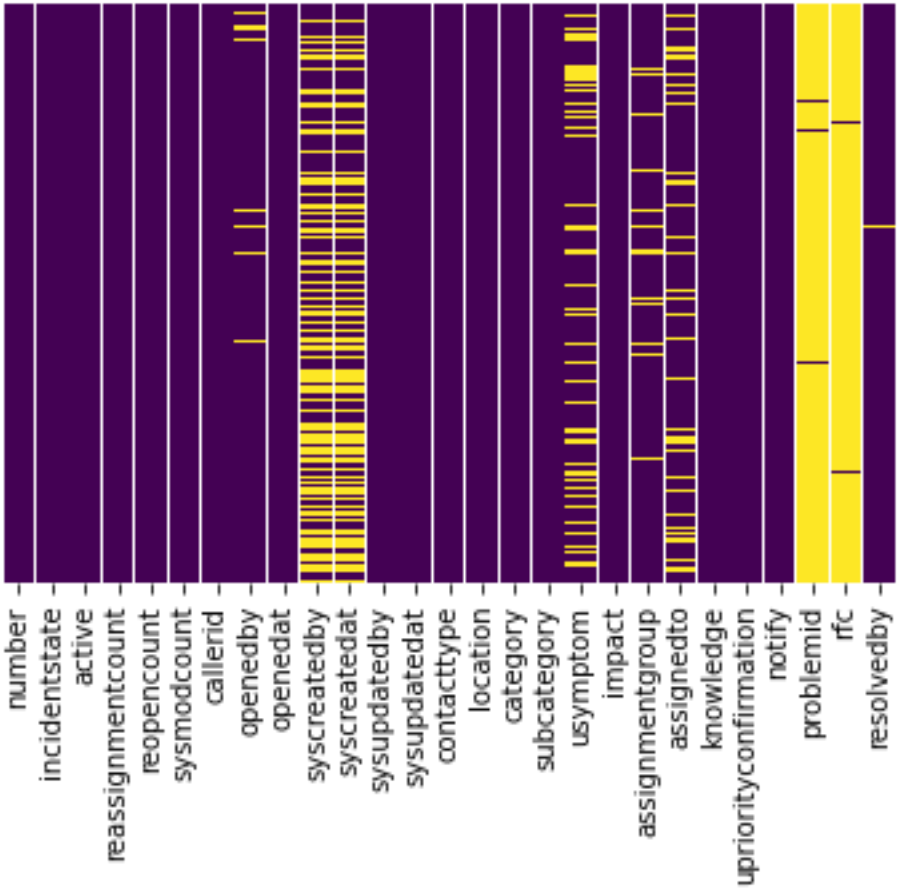
Sr. no		DESCRIPTION
1	No. of rows and columns	141712 rows & 25 columns
2	No. of Missing values	Some rows contain '?' which was replaced by Nan. callerid-29,opened by-4835,syscreatedby-53076,syscreateddat-53076,location-76,subcategory-111,Usymptoms-32964,assignedto-27496,problemid-139417,rfc- 140721,causedby-141689
3	Numerical Variables	Caller_id,opened_by,sys_created_by,sys_updated_by,location,category,subcategory, u_symptom,impact,urgency_priority,assignment_group,assigned_to, problem_id,closed_code,resolved_by
4	Categorical Variables	Incident_state,active,made_sla,impact,urgency,prioruty, knowledge,u_priority_confirma tion,notify
5	Datatypes	bool(3), datetime64[ns](2), int64(3), object(17)
6	Drop unnecessary Columns	Problem_id,caused_by,rfc
7	Remove strings	Columns which contained such as opened by, created by, resolved by-these strings where removed and only numerical values where kept.

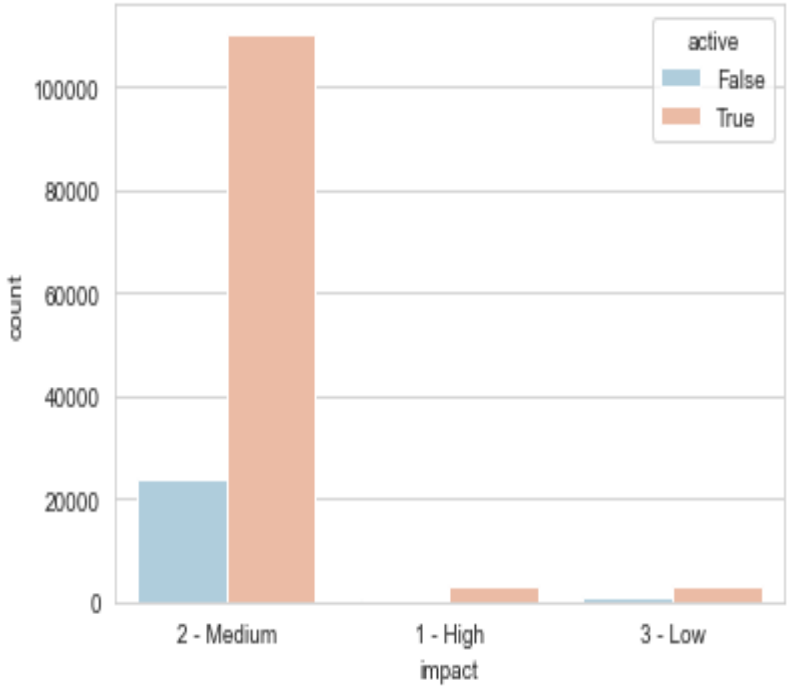
Exploratory Data Analysis & Feature Engineering

Exploratory Data Analysis (EDA)

EXCELR
Raising Excellence

We can use seaborn to create a simple heatmap to see where missing data. and here, we visualize the data impact wise such as High, Medium, Low We visualizes the missing data using heatmap and also visulizes active and unactive users impactwise.

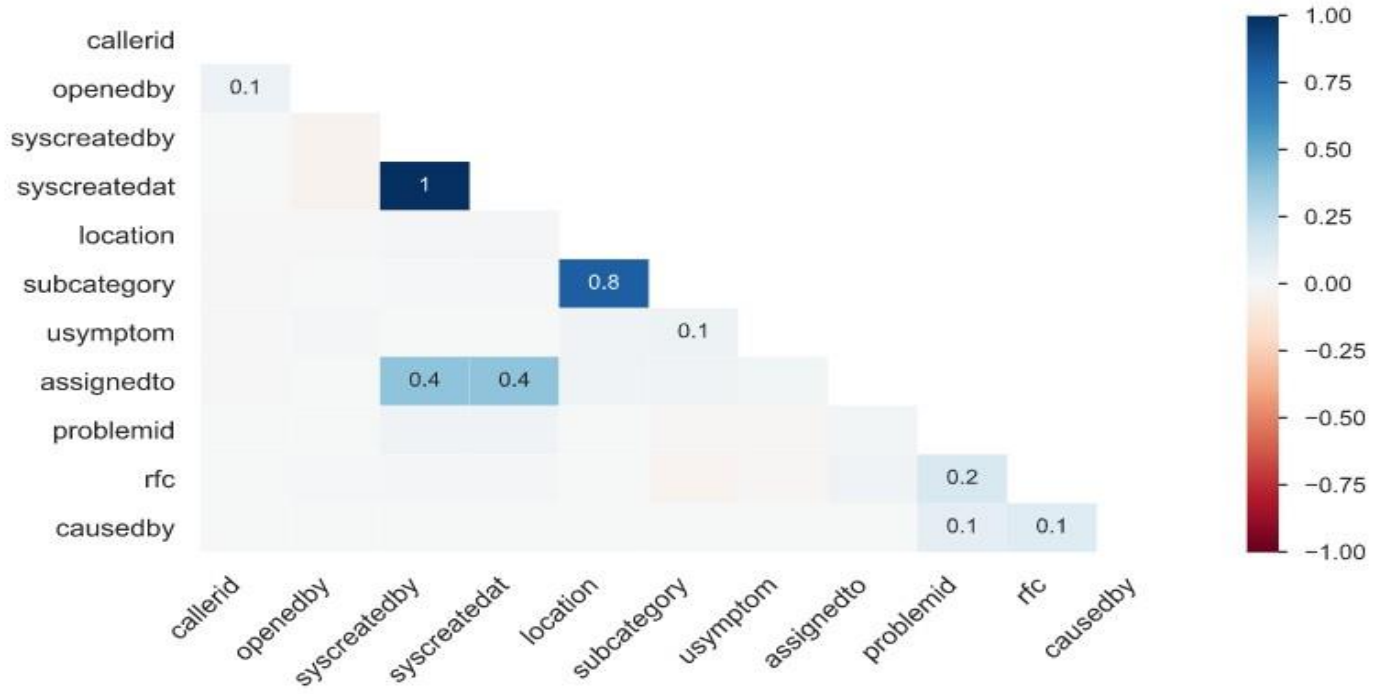




Data Visualization

EXCELR
Raising Excellence

As per the visualization of missing data we can say that,
The proportion of created by and created at missing is likely small enough for reasonable replacement with some form of imputation. looking at column user symptom AND assigned to less amount of data is missing nearly around 10% do data. Looking at the problem id , rfc and caused by it looks like we are just missing too much of that data nearly 99% of data to do something useful with at a basic level.

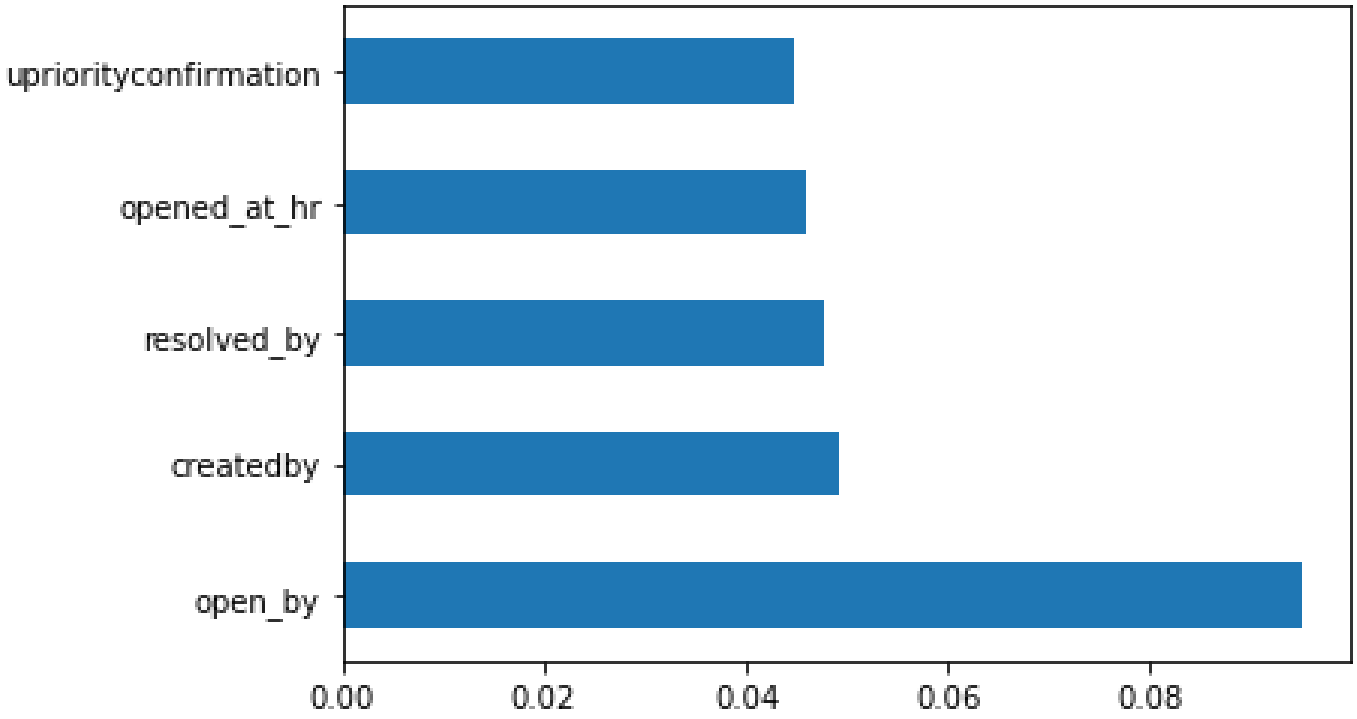


Feature Engineering

EXCELR
Raising Excellence

(A) Feature selection using Extra tree Classifier

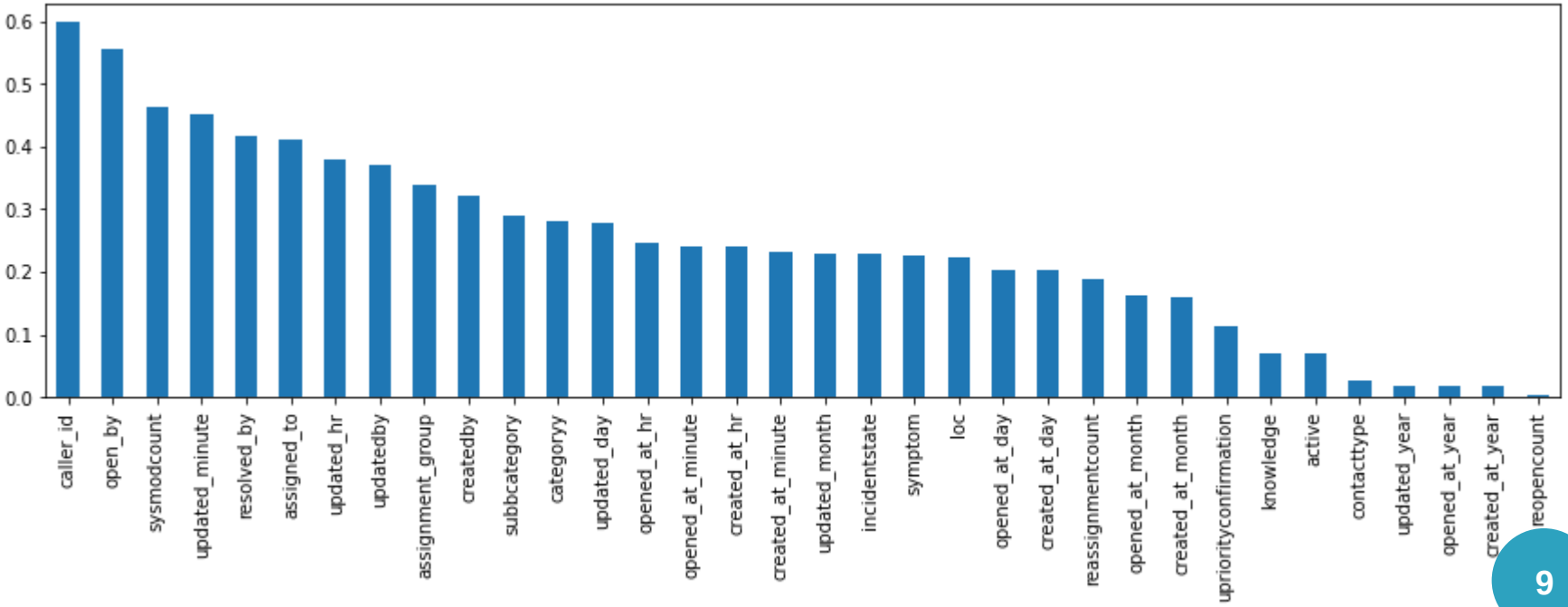
The purpose of the extra tree classifier is to fit a number of randomized decision trees to the data, and in this regard is a from of ensemble learning. Particularly, random splits of all observations.



Feature Engineering

(B) Feature selection using Mutual Information

Mutual information has been successfully adopted in filter feature-selection methods to assess both the relevancy of a subset of features in predicting the target variable and the redundancy with respect to other variables.



Feature Engineering

EXCELR
Raising Excellence

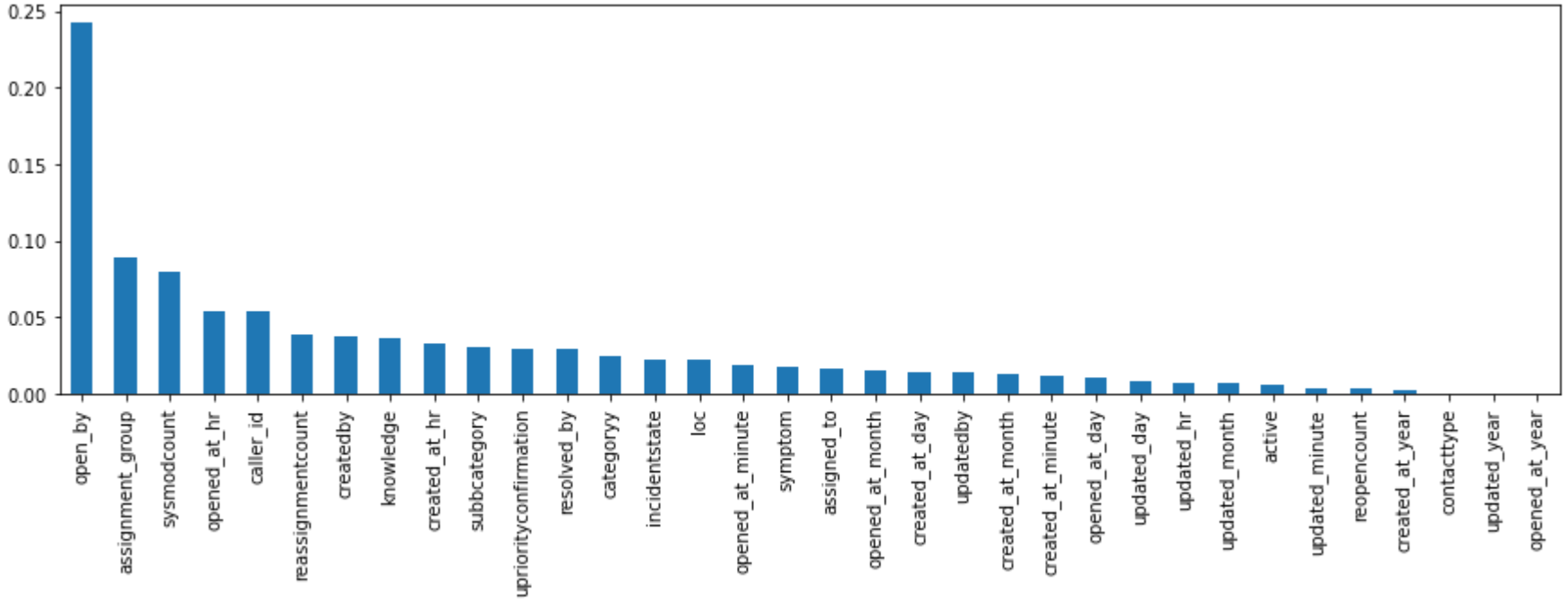
(B) Feature importance using Decision Tree Classifier

open_by	0.242338
assignment_group	0.089101
sysmodcount	0.079730
opened_at_hr	0.054005
caller_id	0.053689
reassignmentcount	0.039402
createdby	0.037564
knowledge	0.036178
created_at_hr	0.032845
subbcategory	0.030735
uprioritycconfirmation	0.029824
resolved_by	0.029232
categoryy	0.024447
incidentstate	0.022956
loc	0.022899
opened_at_minute	0.019180
symptom	0.018262
assigned_to	0.016747
opened_at_month	0.015202
created_at_day	0.014683
updatedby	0.013975
created_at_month	0.013244
created_at_minute	0.011964
opened_at_day	0.011285
updated_day	0.008284

Feature Engineering

EXCELR
Raising Excellence

Feature importance using the Decision Tree Classifier :



Feature Engineering

EXCELR
Raising Excellence

Feature Selection using Chi-Square :

These are top ten features useful for predicting the impact :

	scores	0
8	1.595930e+06	caller_id
9	8.731750e+05	open_by
11	3.492463e+05	updatedby
10	2.530590e+05	createdby
15	7.835950e+04	assigned_to
17	6.064706e+04	resolved_by
12	4.963523e+04	loc
4	2.616341e+04	sysmodcount
13	2.569817e+04	subbcategory
24	2.258494e+04	opened_at_day

- caller Id
- open_by
- updated_by
- createdby
- assigned_to
- resolved_by
- loc
- sysmodcount
- subbcategory
- opened_at_day

Model Building

Predicting Model Accuracy :

EXCELR
Raising Excellence

Random forest

0.9843812480888199

	precision	recall	f1-score	support
1	0.81	0.91	0.86	1098
2	1.00	0.99	0.99	40176
3	0.79	0.97	0.87	1239
accuracy			0.98	42513
macro avg	0.87	0.95	0.91	42513
weighted avg	0.99	0.98	0.98	42513

Decision tree

0.983746148237009 9

	precision	recall	f1-score	support
1	0.78	0.92	0.84	1098
2	1.00	0.99	0.99	40176
3	0.80	0.96	0.87	1239
accuracy			0.98	42513
macro avg	0.86	0.96	0.90	42513
weighted avg	0.99	0.98	0.98	42513

KNeighboursClassifier

0.907675299320208

	precision	recall	f1-score	support
1	0.32	0.97	0.48	1098
2	1.00	0.90	0.95	40176
3	0.42	0.96	0.59	1239
accuracy			0.91	42513
macro avg	0.58	0.94	0.67	42513
weighted avg	0.96	0.91	0.93	42513

XG Boost

0.957354221061792 9

	precision	recall	f1-score	support
1	0.57	0.85	0.68	1098
2	0.99	0.96	0.98	40176
3	0.56	0.91	0.70	1239
accuracy			0.96	42513
macro avg	0.71	0.91	0.79	42513
weighted avg	0.97	0.96	0.96	42513

Predicting Model Accuracy :

EXCELR
Raising Excellence

MPL classifier

0.669856881424505

	precision	recall	f1-score	support
1	0.08	0.76	0.15	1098
2	0.99	0.66	0.79	40176
3	0.18	0.80	0.29	1239
accuracy			0.67	42513
macro avg	0.42	0.74	0.41	42513
weighted avg	0.94	0.67	0.76	42513

Naive bayes

0.4940371180580576

	precision	recall	f1-score	support
1	0.04	0.41	0.07	1098
2	0.96	0.50	0.66	40176
3	0.05	0.41	0.09	1239
accuracy			0.49	42513
macro avg	0.35	0.44	0.27	42513
weighted avg	0.91	0.49	0.62	42513

spred_gnb

0.4940371180580576

	precision	recall	f1-score	support
1	0.04	0.41	0.07	1098
2	0.96	0.50	0.66	40176
3	0.05	0.41	0.09	1239
accuracy			0.49	42513
macro avg	0.35	0.44	0.27	42513
weighted avg	0.91	0.49	0.62	42513

spred_mnb

0.4971420506668548 5

	precision	recall	f1-score	support
1	0.04	0.14	0.06	1098
2	0.96	0.50	0.66	40176
3	0.04	0.58	0.08	1239
accuracy			0.50	42513
macro avg	0.35	0.41	0.27	42513
weighted avg	0.91	0.50	0.63	42513

Final Model : Random Forest

Algorithm_name : RandomForestClassifier

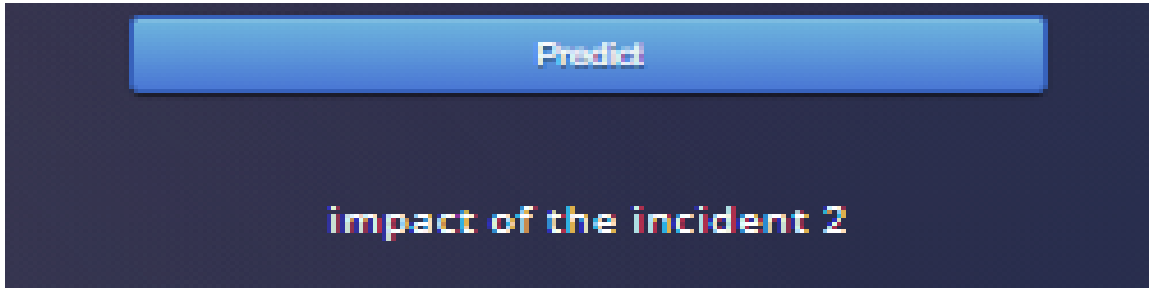
Accuracy : 0.9843812480888199

	precision	recall	f1-score	support
1	0.81	0.91	0.86	1098
2	1.00	0.99	0.99	40176
3	0.79	0.97	0.87	1239
accuracy			0.98	42513
macro avg	0.87	0.95	0.91	42513
weighted avg	0.99	0.98	0.98	42513

Sr.no	Model	Accuracy
1	model_rf	98.438125
2	model_dtree	98.374615
3	model_knn	90.767530
4	model_xgb	95.735422
5	model_mlp	88.133042
6	spred_gnb	49.403712
7	spred_mnb	49.714205

Model Deployment using Flask

Model Deployment: Input values & Output



Challenges faced ?

1. Feature Selection
2. How to improve f1 score & accuracy of model
3. Deployment

Thank you