

# Incident Impact Prediction

**Group: 4**

**Dhanya GD**

**Macchindra Ganpat Shinde**

**Vikas Bhagwan Chaudhari**

**Rahul Kumar**

**Sourajit Dey**

**Nithish Kumar CV**

**Mentor: Parth Sagar**

**Date: 10/02/2022**

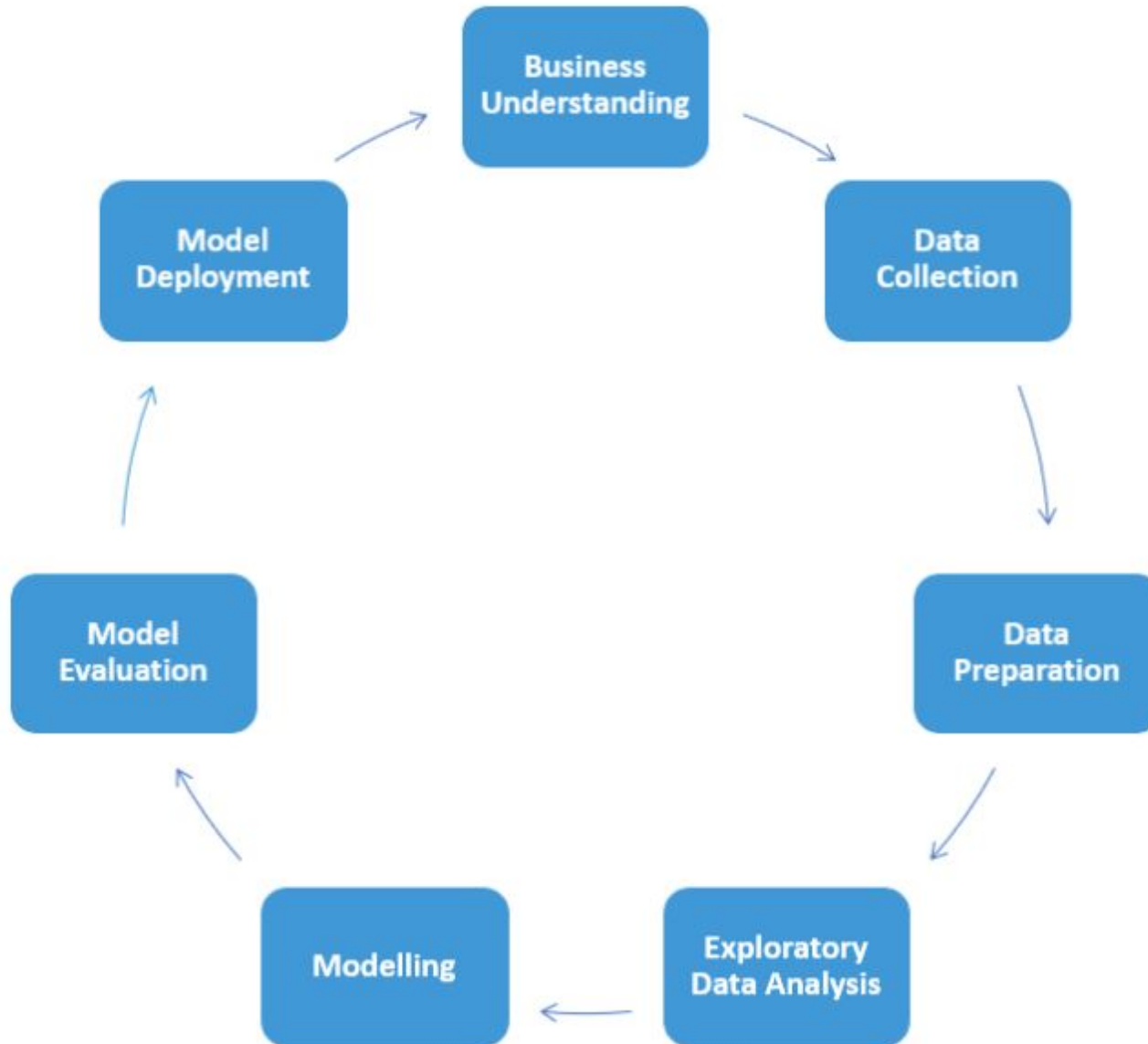
# Business Problem:

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

## Objective:

- Whenever a service being provided to the customer is disrupted, the customer raises a complaint to the customer service; who in turn raises a ticket.
- When an incident is logged, a lot of data related to the incident is recorded for ex: time, status ...
- Our objective is to predict the impact of an incident i.e whether it is High, Medium or Low. So that the company can optimize the resource allocation to focus on the important incidents.
- A quicker resolution translates to customer satisfaction/retention which leads to better business.

# Project Architecture



# Project Flow

- Understanding the business problem.
- Data collection
- Data preparation.
- EDA and Feature Engineering
- Model building
- Model training and evaluation
- Deployment

# **Exploratory Data Analysis (EDA)**

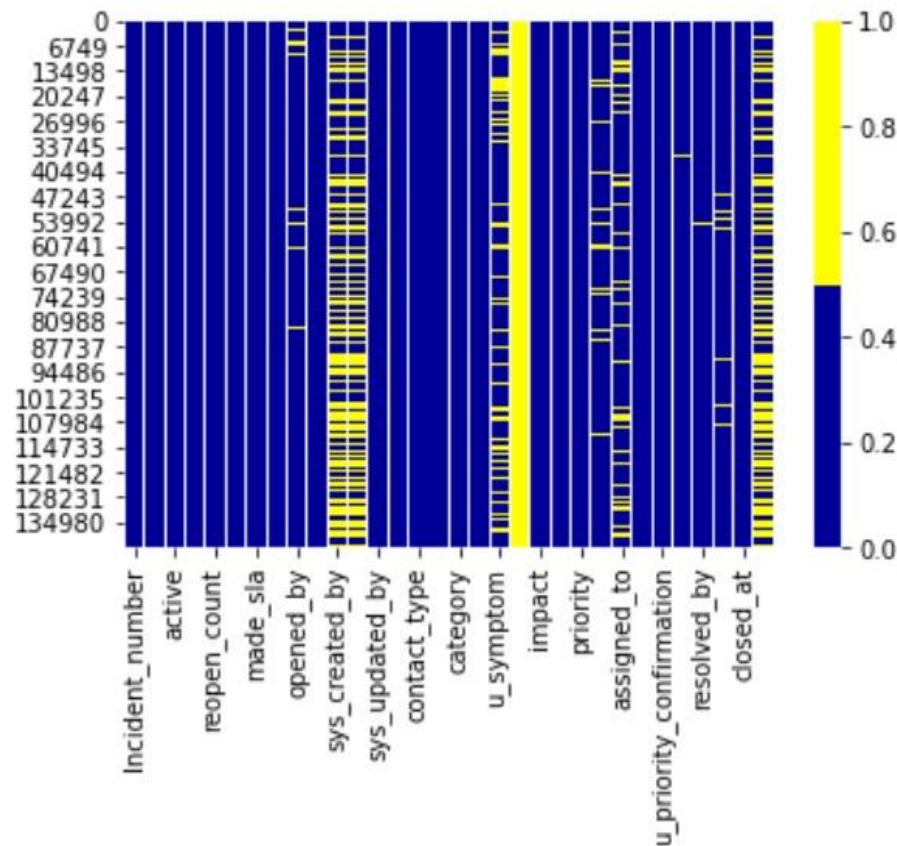
## Dataset details:

- The dataset has 141712 records, 36 columns with 'impact' column as **target**.

column type	No of columns
Numeric columns	3
Boolean columns	5
Categorical columns	23
Datetime columns	5
<b>Total</b>	<b>36</b>

## Null values:

- Null values recorded as '?' (mostly) and '-100'.
- Columns with more than 98% missing values: 'cmdb\_ci', 'problem\_id', 'rfc', 'vendor', 'caused\_by'.
- 'sys\_created\_by' and 'sys\_created\_at' have 37% null values.



**Null values by column:**

col_name	null_vals	% percentage of null vals
opened_by	4835	3
sys_created_by	53076	37
sys_created_at	53076	37
u_symptom	32964	23
cmdb_ci	141267	100
assignment_group	14213	10
assigned_to	27496	19
problem_id	139417	98
rfc	140721	99
vendor	141468	100
caused_by	141689	100
closed_code	714	1
resolved_at	3141	2



## Unique values:

- Mix of high and low cardinality columns.

Unique values < 100

	columns	num_unique_vals
1	incident_state	8
5	contact_type	5
7	category	59
10	impact	3
11	urgency	3
12	priority	4
13	assignment_group	79
15	closed_code	17

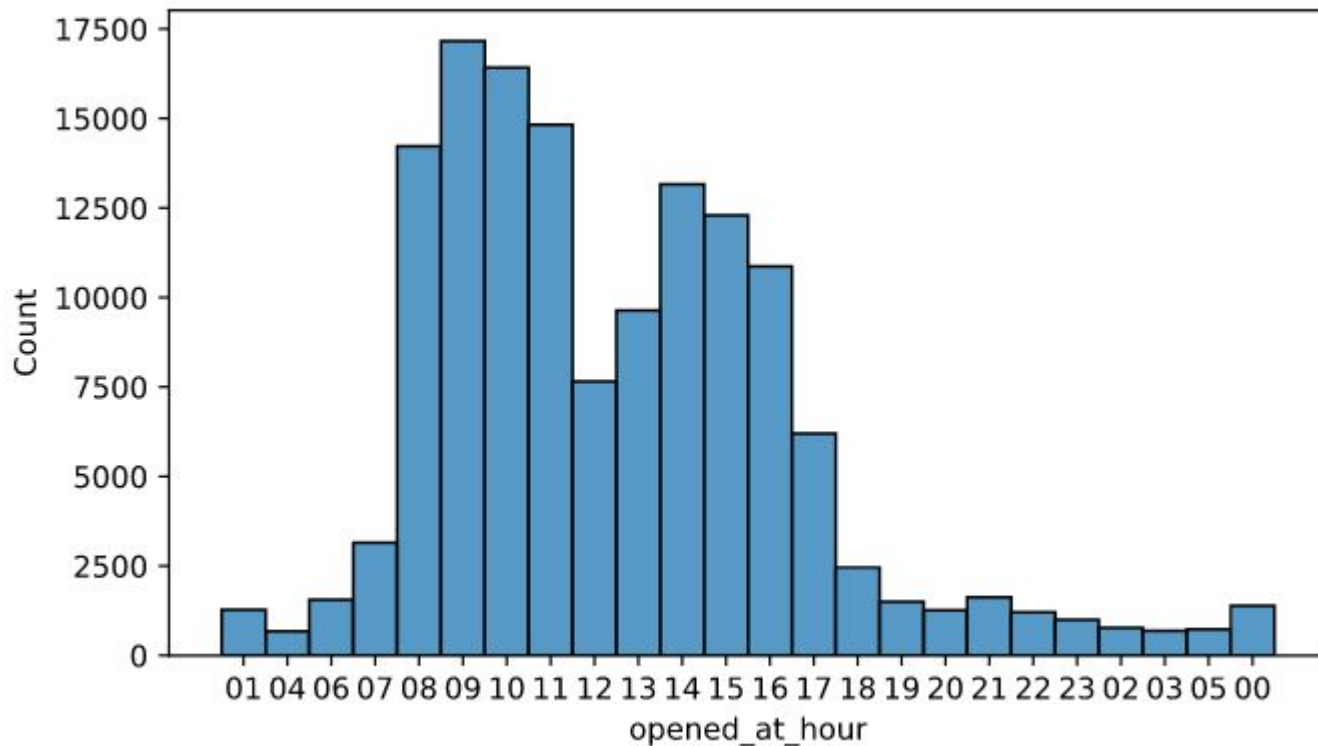
Unique values > 100

	columns	num_unique_vals
0	number	24918
2	caller_id	5245
3	opened_by	208
4	sys_updated_by	846
6	location	225
8	subcategory	255
9	u_symptom	526
14	assigned_to	235
16	resolved_by	217

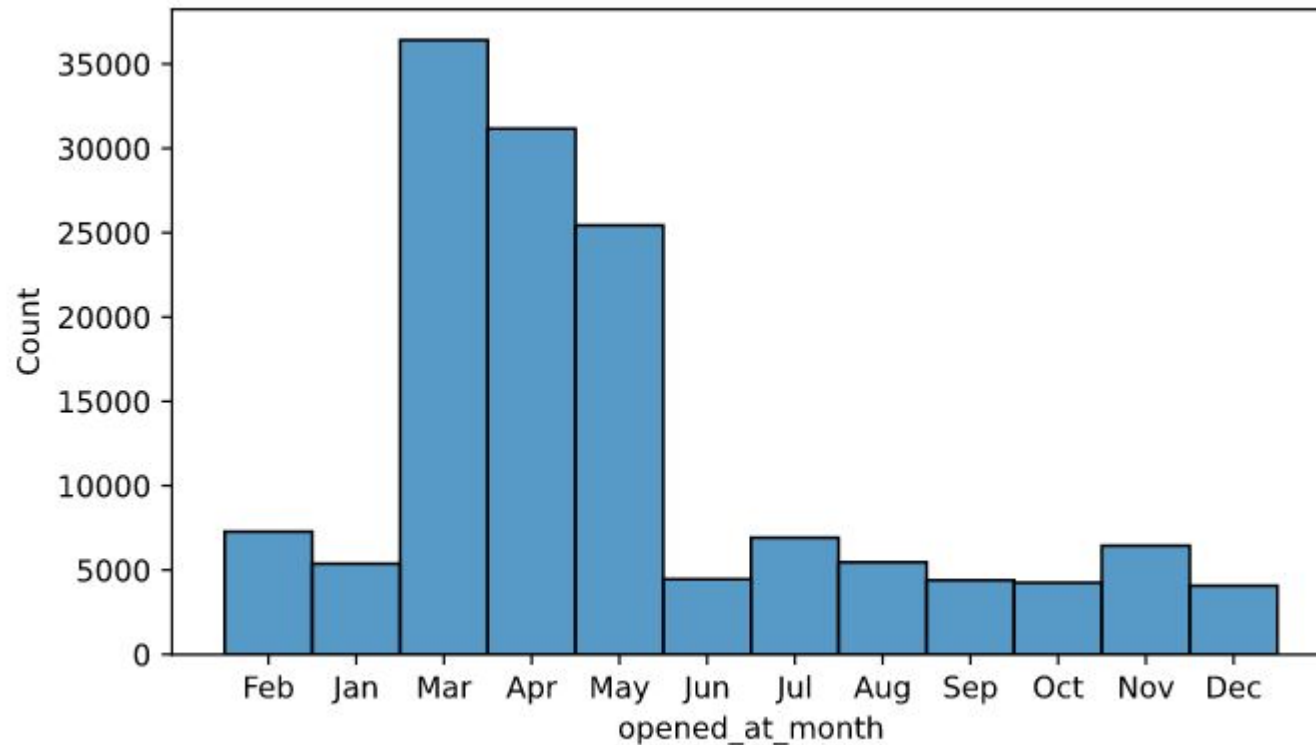
## **Assumptions / Queries:**

- Date and time columns may affect our target variable.
- 'incident no' may not affect target.
- 'opened\_by' may not affect the target.
- 'opened at' - should be investigated further.
- Does location of incident affect the impact column?
- Does contact type (phone, email etc.) affect impact column?
- Does 'u\_symptom affect' the impact column?
- Having a prior knowledge of the incident may lead to quicker resolution and may have a relation to impact.

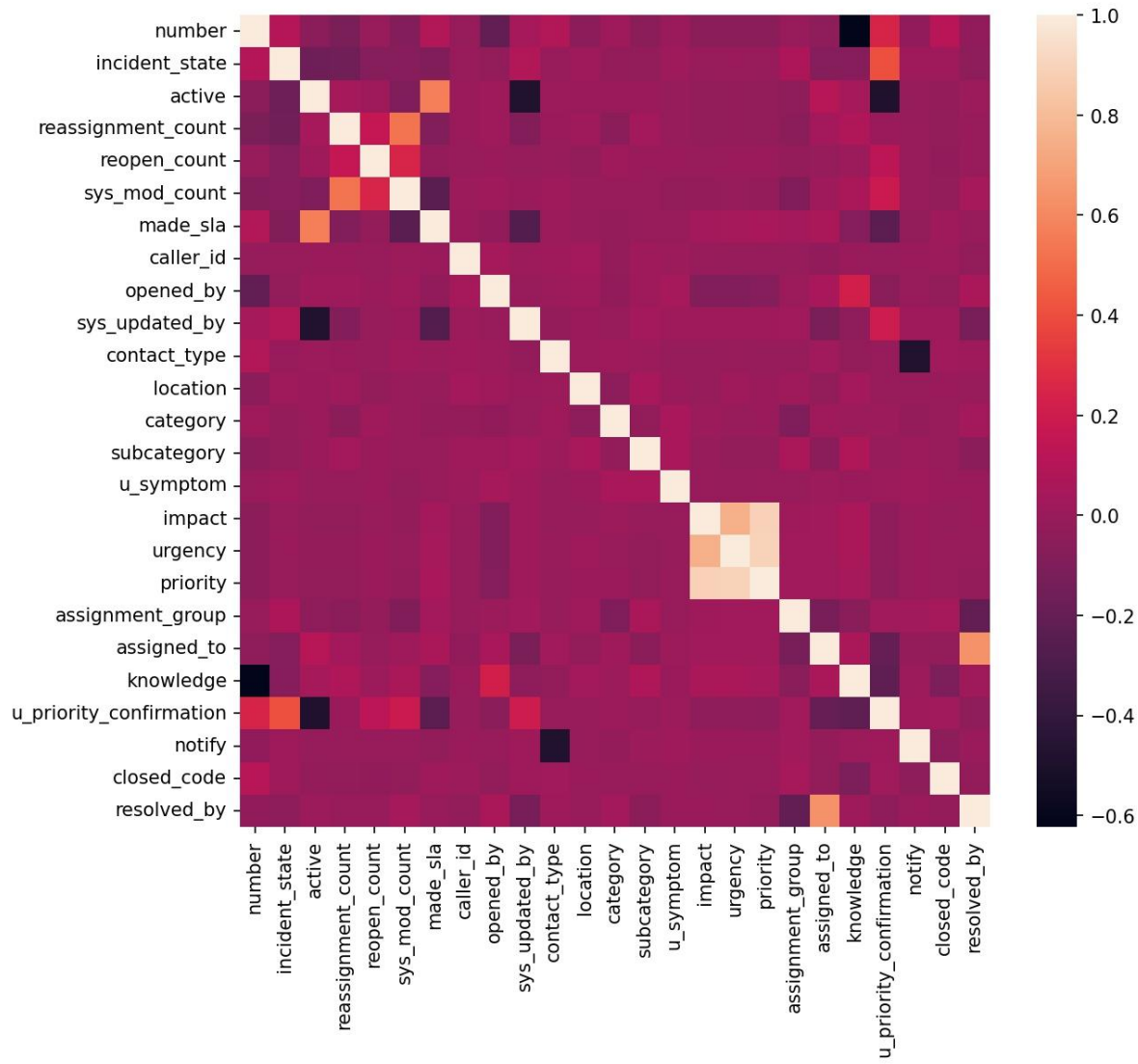
On splitting the 'opened\_at' column into date and time, we observe that most of the incidents get reported during 8:00 hrs to 17:00 hrs



On splitting the 'opened\_at' column into months, we observe that most of the incidents get reported during during months of march , april, and may.

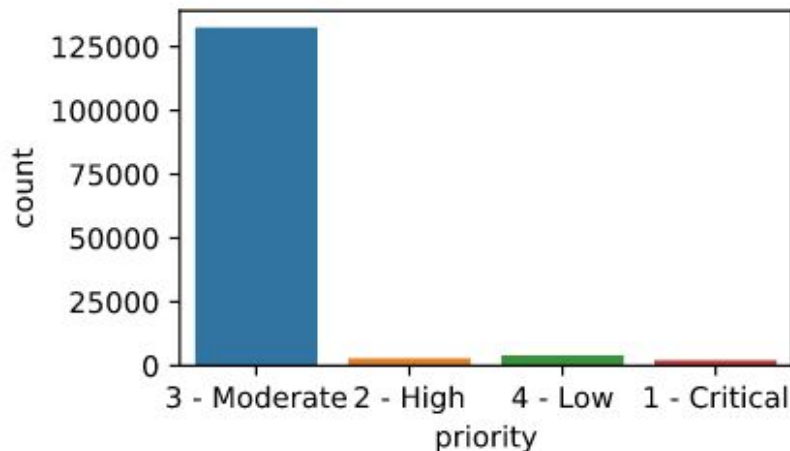
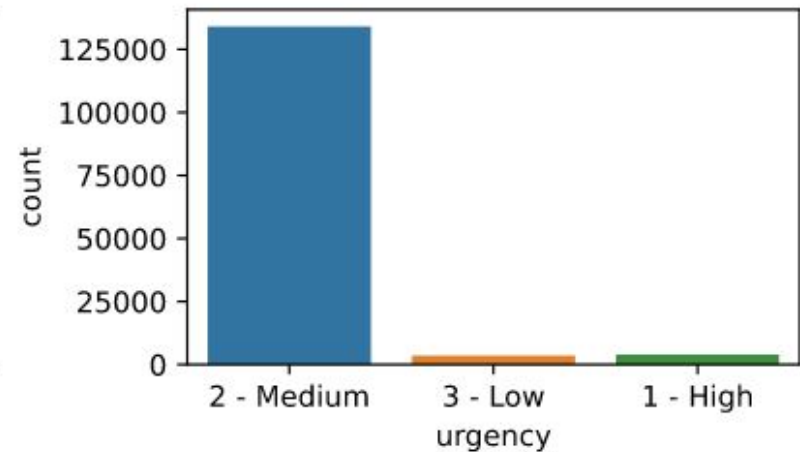
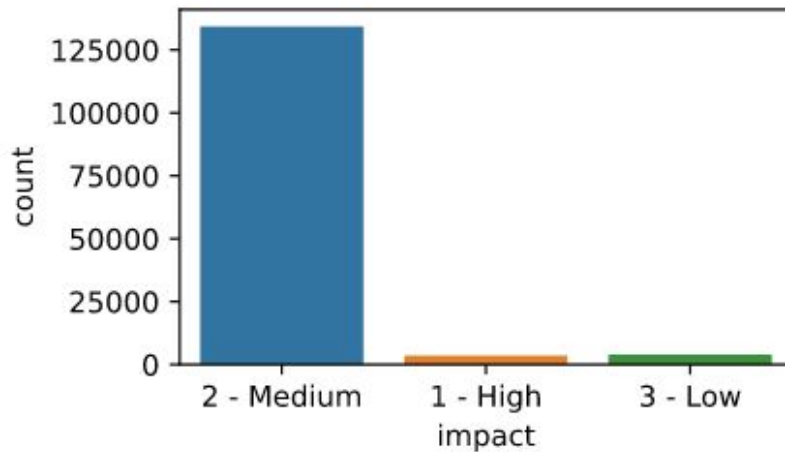


- From the heatmap we see that except for 'urgency' and 'priority' none of the other features have a strong correlation with the target.
- There are few strong correlations between feature columns.



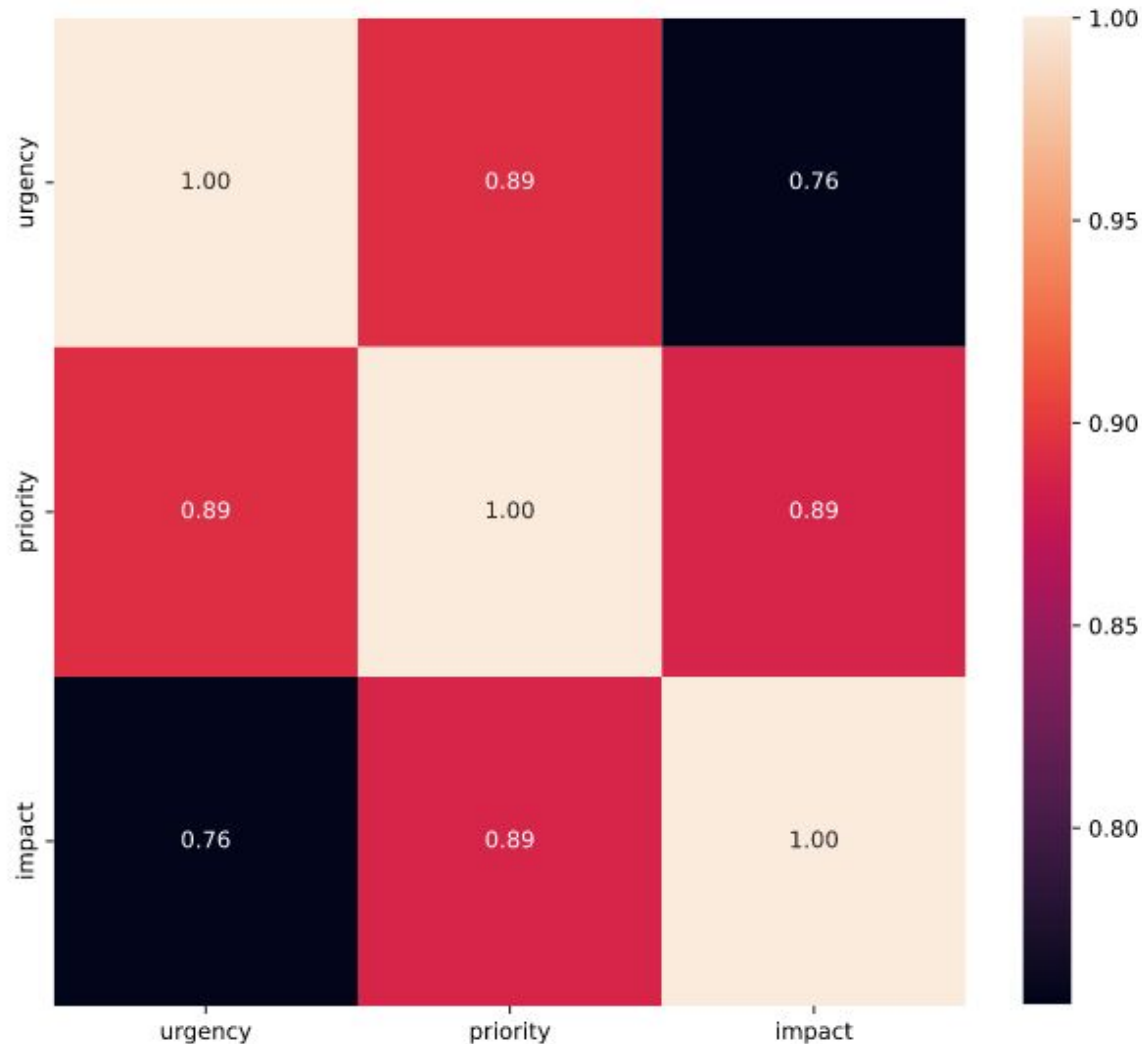
We observe that the target distribution is imbalanced with most of the incidents classified as having a 'Medium' impact. The columns 'urgency' and 'priority' follow a similar trend.

**Distribution of target('impact') compared to similar features**



## Relationship between 'urgency', 'Priority' and 'Impact':

The above columns are very highly correlated with 'impact'. Will these two features play a dominant role in predictions?

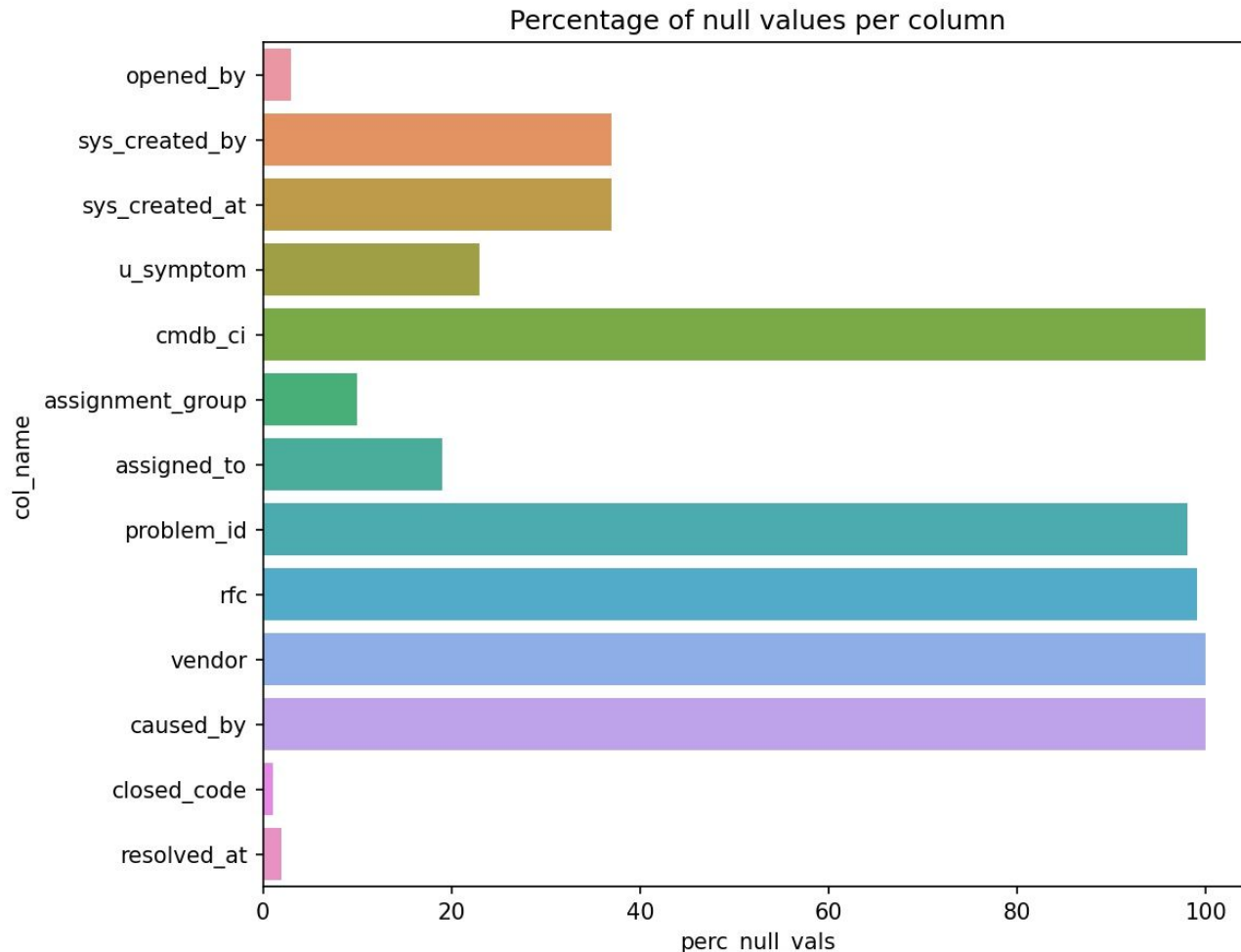


# Feature Engineering

## Drop the columns with most null values:

The following columns have more than 35% null values and can be dropped:

'cmdb\_ci', 'problem\_id', 'rfc', 'vendor', 'caused\_by',  
'sys\_created\_by', 'sys\_created\_at'





## Imputing Null values:

The following columns have fewer null values and will be imputed as described below.

COLUMN NAME	IMPUTED VALUE	IMPUTATION STRATEGY
'caller_id'	Mode	Mode
'opened_by'	Mode	Mode
'location'	Mode	Mode
'category'	Mode	Mode
'subcategory'	Mode	Mode
'u_symptom'	Mode	Mode
'assignment_group'	Mode	Mode
'assigned_to'	Mode	Mode
'resolved_by'	Mode	Mode
'incident_state'	Mode	Mode
'closed_code'	-	padding

## Encoding Categorical columns:

Columns with boolean categories converted to 0 and 1.

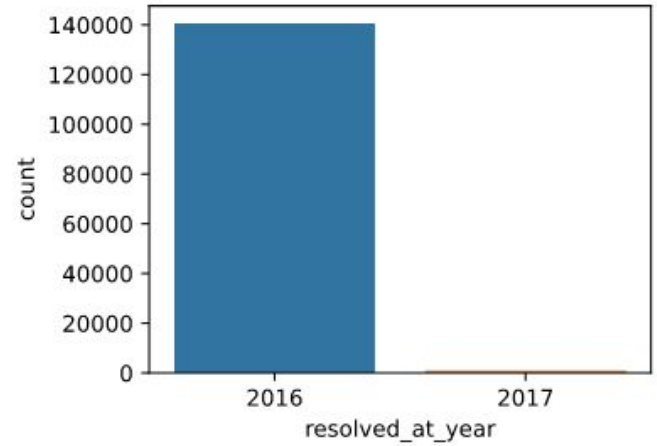
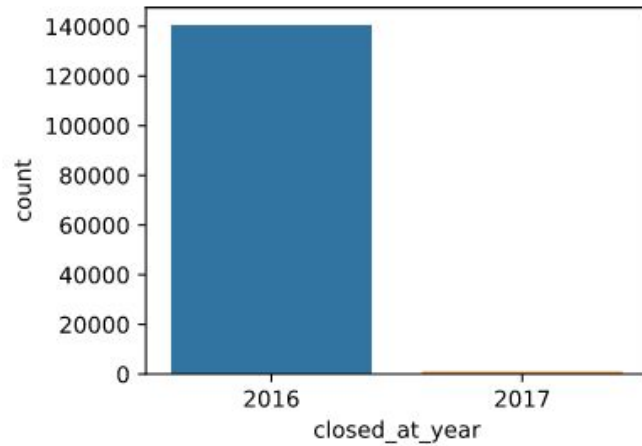
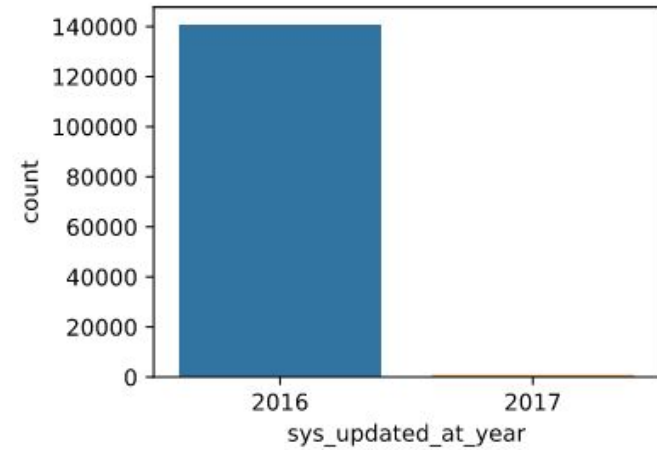
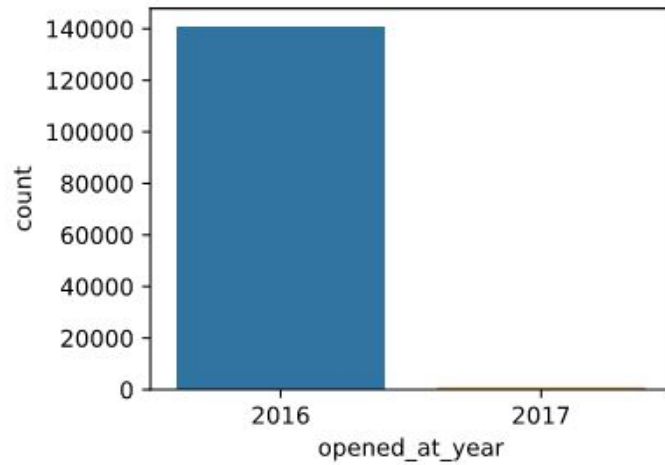
COLUMN	LABELS	BOOL CATEGORY
'active'	TRUE, FALSE	TRUE-1, FALSE-0
'made_sla'	TRUE, FALSE	TRUE-1, FALSE-0
'knowledge'	TRUE, FALSE	TRUE-1, FALSE-0
'u_priority_confirmation'	TRUE, FALSE	TRUE-1, FALSE-0
'notify'	'Send Email', Do Not Notify	Do Not Notify-0, 'Send Email'-1

Columns with more than two labels encoded using ordinal encoder.

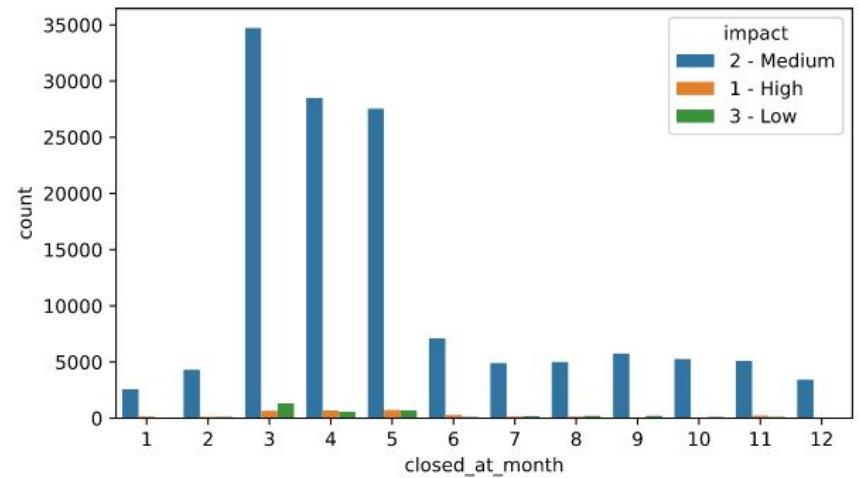
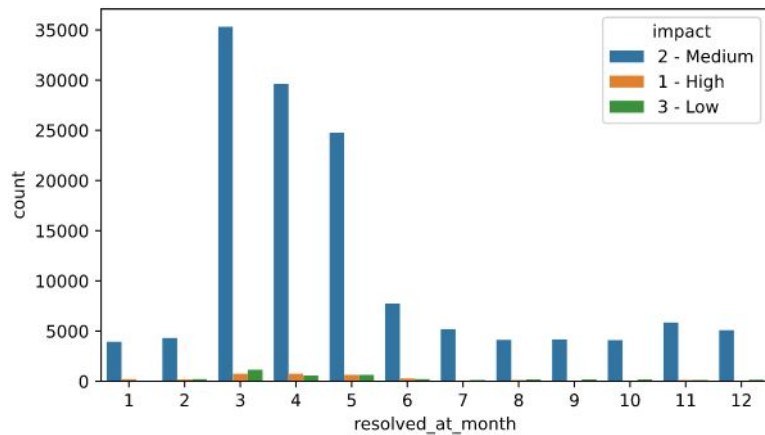
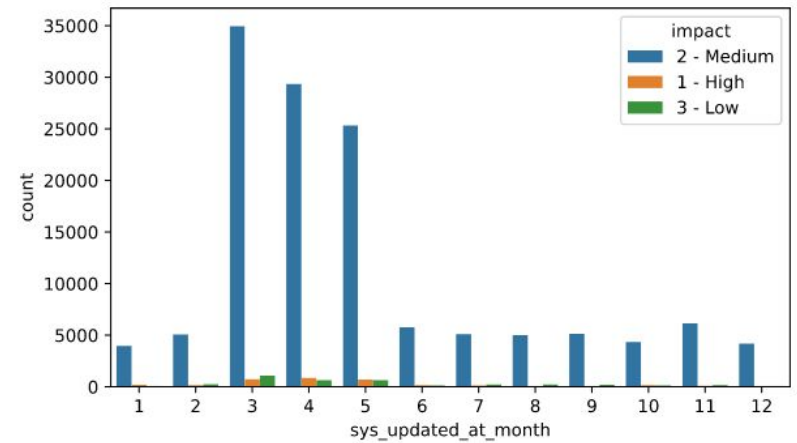
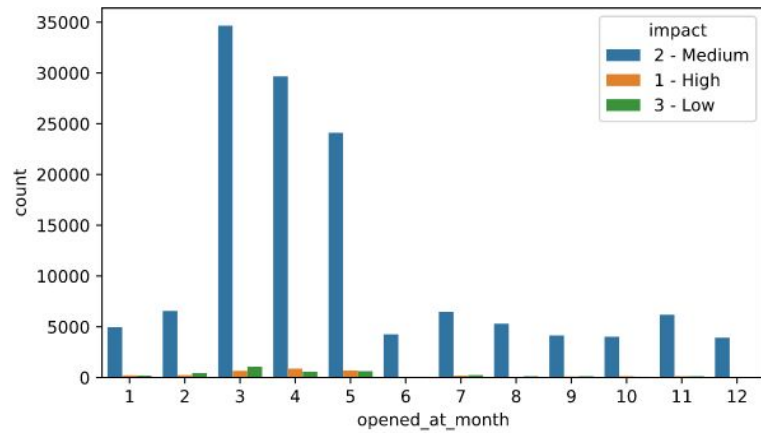
## **Extracted features from columns having datetime data**

Features like 'Year' , 'Month', 'Week', 'Day', 'Hour', 'Minute' were extracted from the datetime features.

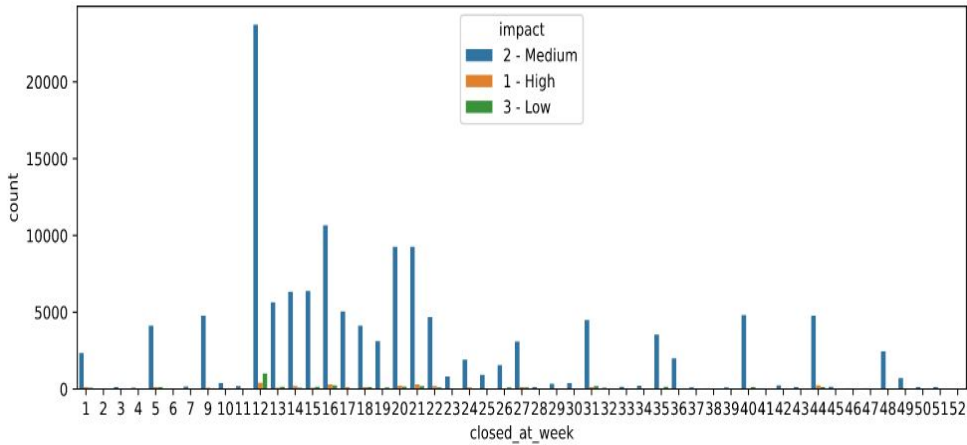
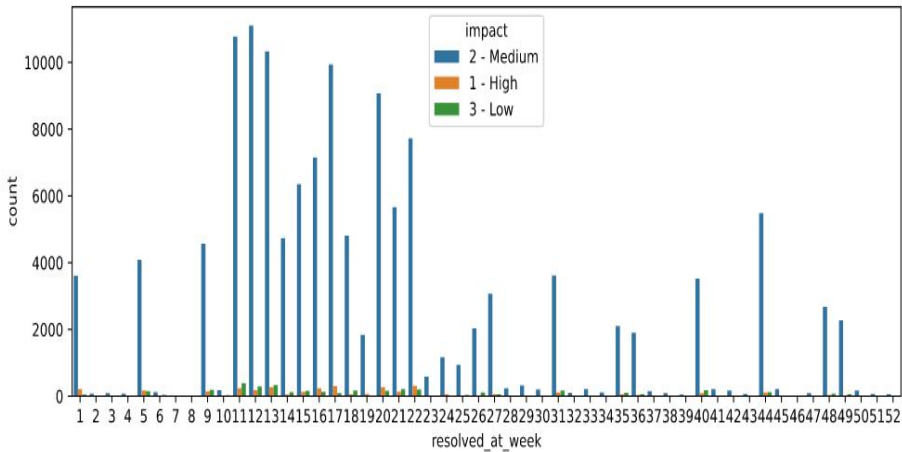
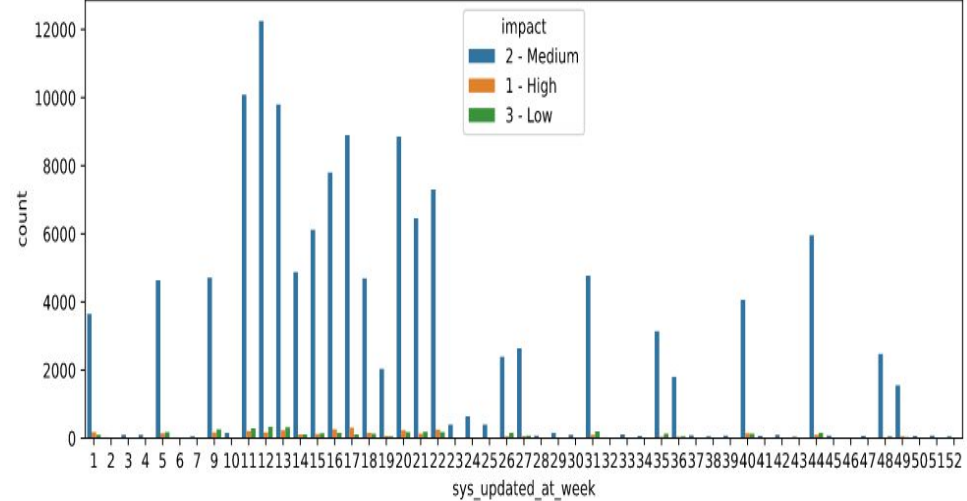
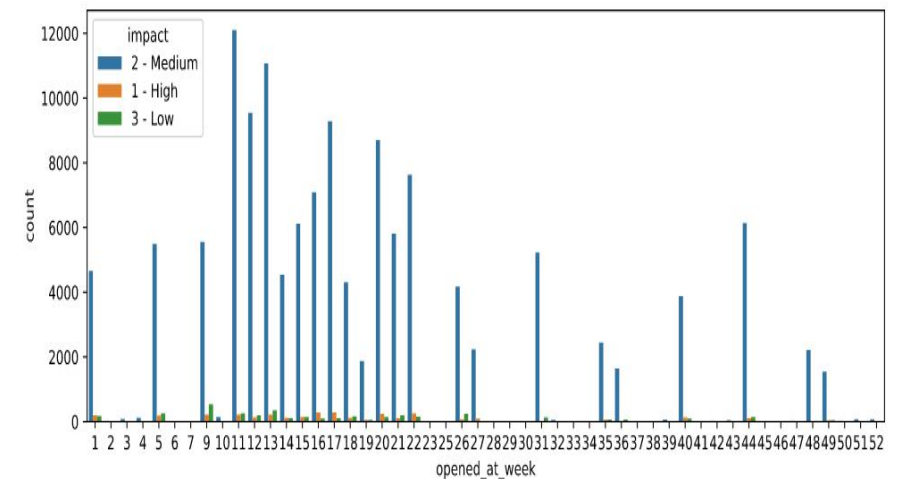
# Year wise extracted features



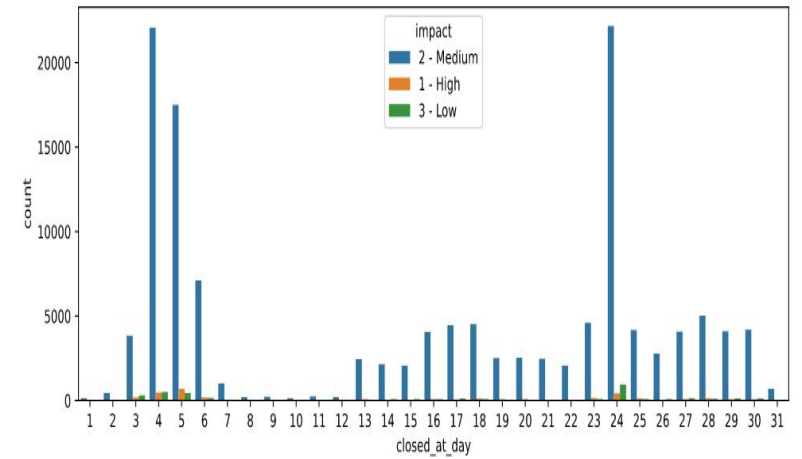
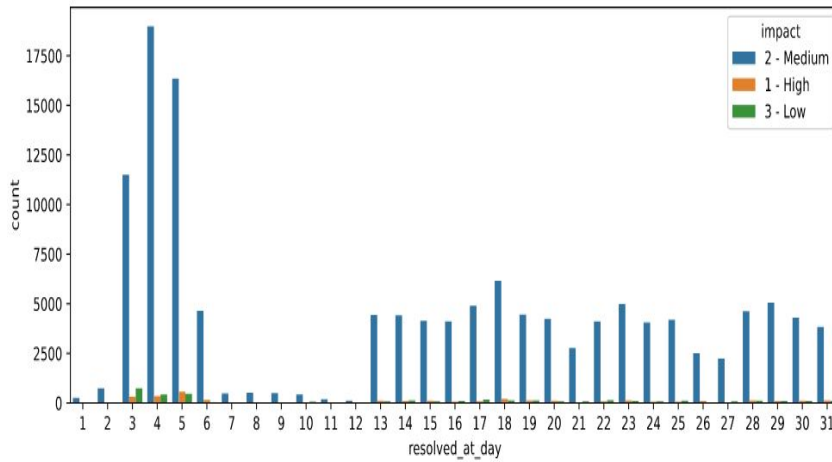
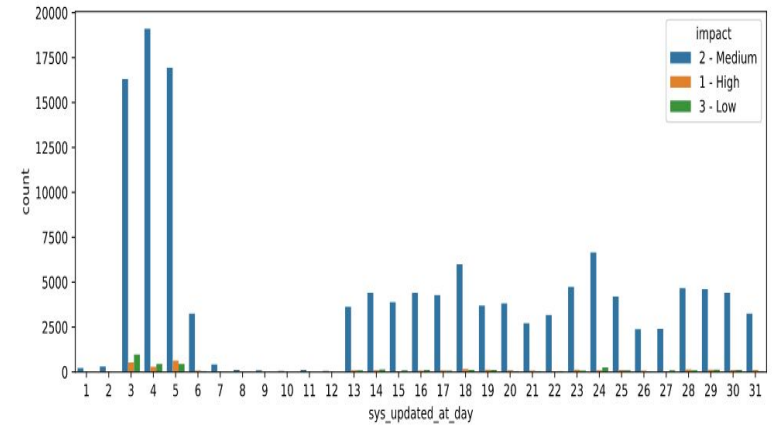
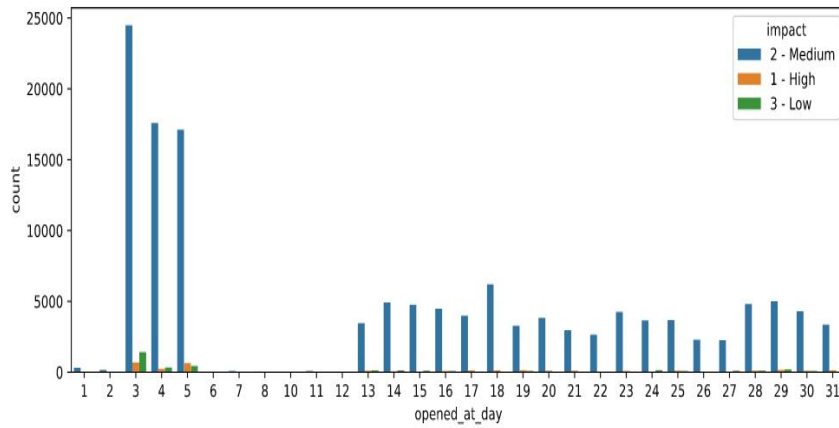
# Month wise extracted features



# Week wise extracted features



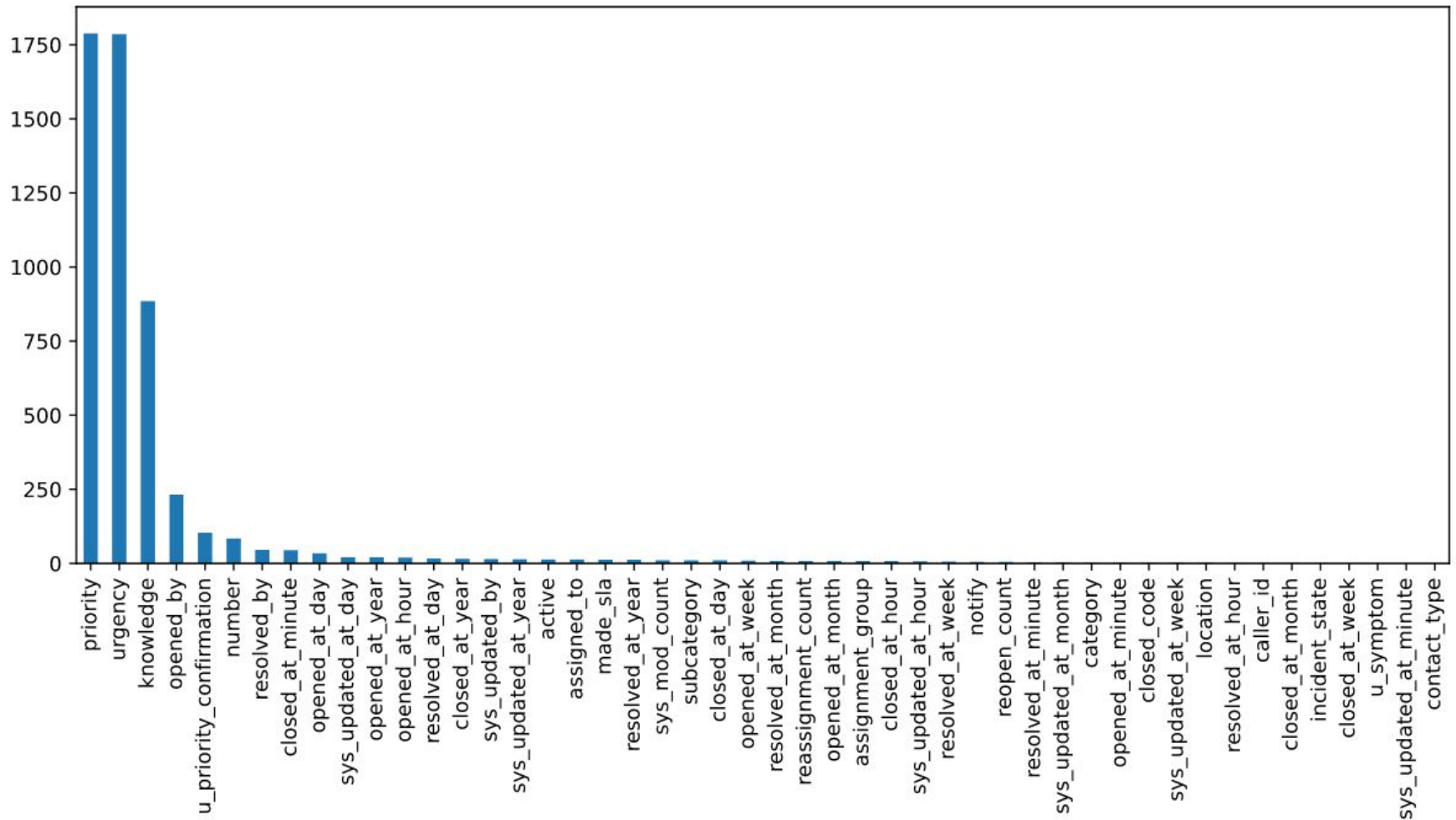
# Day wise extracted features



# Feature selection using Chi-Square and KBest

## Features selected:

'urgency', 'priority', 'knowledge', 'opened\_by', 'u\_priority\_confirmation',  
'number' , 'closed\_at\_minute', 'resolved\_by', 'opened\_at\_day',  
'assigned\_to', 'sys\_updated\_at\_day', 'opened\_at\_hour', 'resolved\_at\_day'





## **Mutual information classification:**

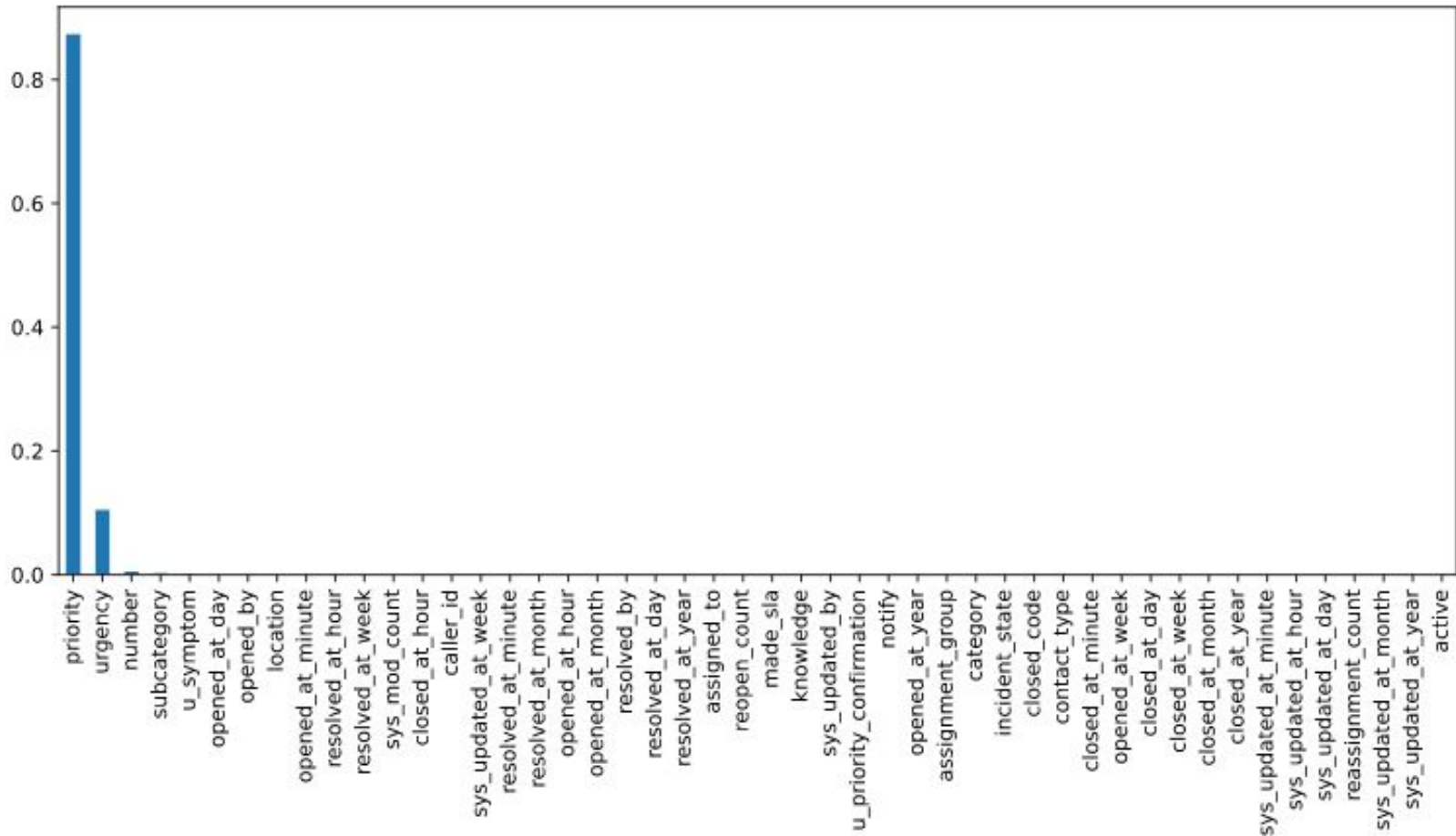
### **Features selected :**

'active', 'made\_sla', 'urgency', 'priority', 'contact\_type', 'category',  
'assignment\_group', 'number', 'caller\_id', 'opened\_by', 'sys\_updated\_by',  
'subcategory', 'u\_symptom', 'assigned\_to', 'resolved\_by'.

# Feature importance using decision trees:

## Features selected :

'priority', 'urgency', 'number', 'subcategory', 'u\_symptom',  
'opened\_at\_day', 'opened\_by', 'location', 'opened\_at\_minute',  
'resolved\_at\_hour', 'resolved\_at\_week', 'sys\_mod\_count', 'closed\_at\_hour'



## list of features removed:.

Dropped features	Reason
cmdb_ci', 'problem_id', 'rfc', 'vendor', 'caused_by', 'sys_created_by', 'sys_created_at'	More than 20% null values
'closed_at'	may not bear any relation to incident impact
'contact_type'	almost all incidents are reported via phone
'sys_mod_count', 'made_sla', 'knowledge', 'u_priority_confirmation', 'notify', 'opened_at_year', 'opened_at_month', 'opened_at_week', 'opened_at_day', 'opened_at_hour', 'opened_at_minute', 'resolved_at_year', 'resolved_at_month', 'resolved_at_week', 'resolved_at_day', 'resolved_at_hour', 'resolved_at_minute', 'sys_updated_at_year', 'sys_updated_at_month', 'sys_updated_at_week', 'sys_updated_at_day', 'sys_updated_at_hour', 'sys_updated_at_minute',	The intersection of feature sets from Chi2, mutual information classification and decision trees excluded these features hence dropped.

## list of features selected:

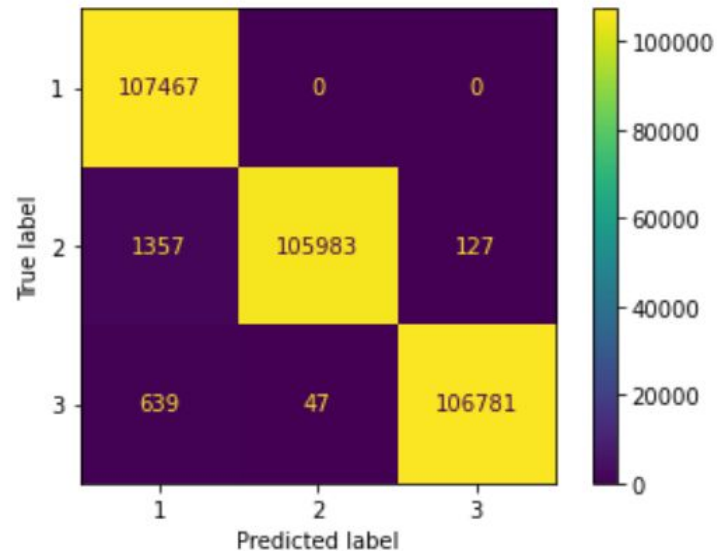
'priority', 'opened\_by', 'number', 'urgency'

# **Model Building**

**Experimenting with different models:**  
Classification reports and confusion matrices

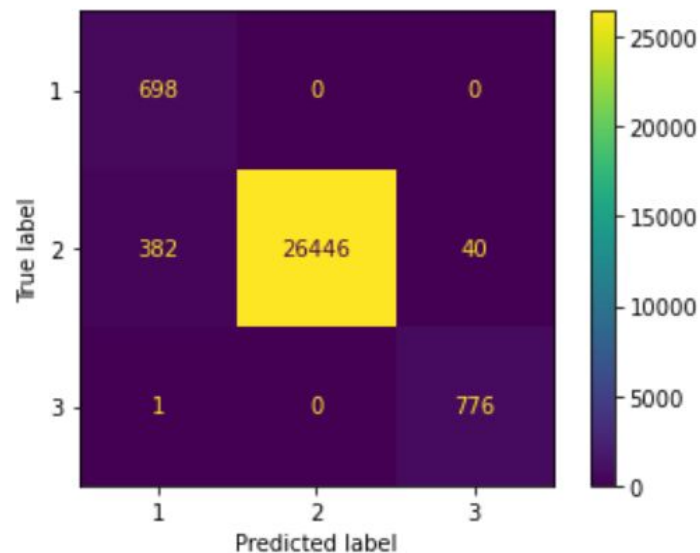
## Gaussian naive bayes: Train data

	precision	recall	f1-score	support
1	0.98	1.00	0.99	107467
2	1.00	0.99	0.99	107467
3	1.00	0.99	1.00	107467
accuracy			0.99	322401
macro avg	0.99	0.99	0.99	322401
weighted avg	0.99	0.99	0.99	322401



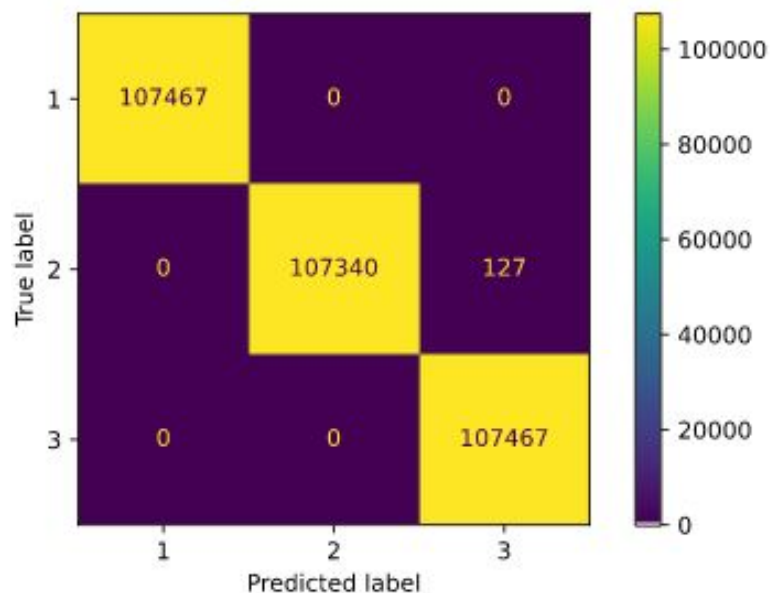
## Gaussian Naive bayes: Test data

	precision	recall	f1-score	support
1	0.65	1.00	0.78	698
2	1.00	0.98	0.99	26868
3	0.95	1.00	0.97	777
accuracy			0.99	28343
macro avg	0.87	0.99	0.92	28343
weighted avg	0.99	0.99	0.99	28343



## Logistic regression: Train data

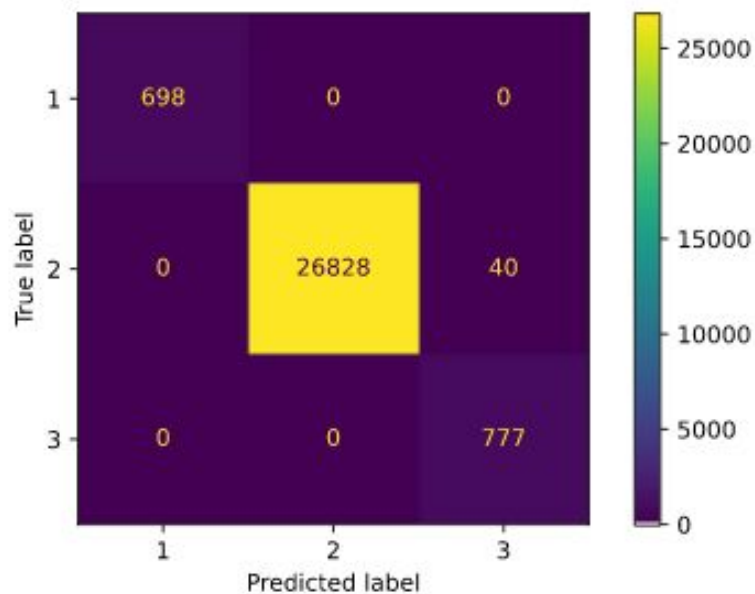
	precision	recall	f1-score	support
1	1.00	1.00	1.00	107467
2	1.00	1.00	1.00	107467
3	1.00	1.00	1.00	107467
accuracy			1.00	322401
macro avg	1.00	1.00	1.00	322401
weighted avg	1.00	1.00	1.00	322401





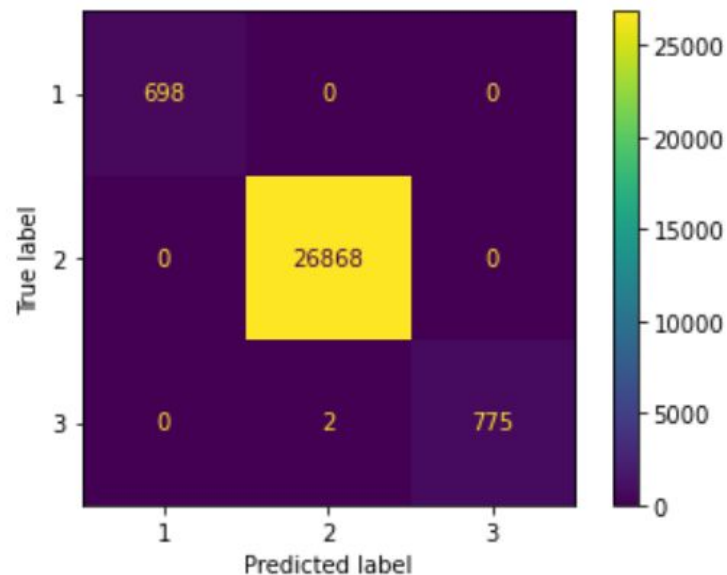
## Logistic regression: Test data

	precision	recall	f1-score	support
1	1.00	1.00	1.00	698
2	1.00	1.00	1.00	26868
3	0.95	1.00	0.97	777
accuracy			1.00	28343
macro avg	0.98	1.00	0.99	28343
weighted avg	1.00	1.00	1.00	28343



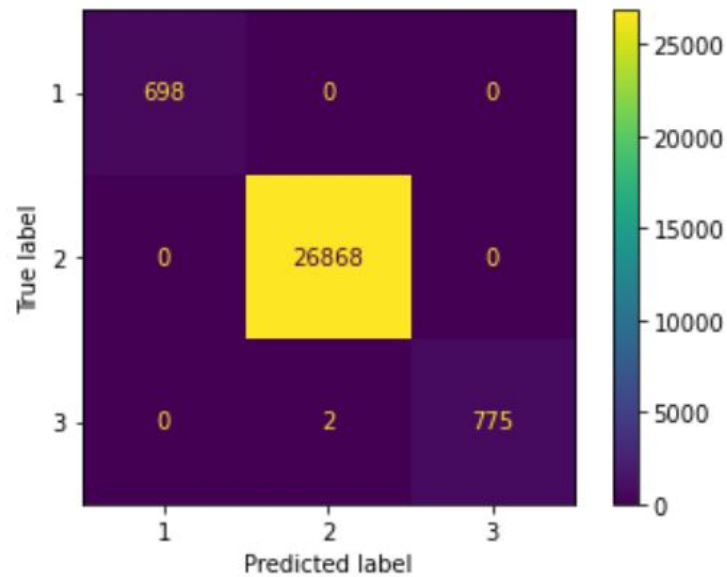
## Decision tree: Train data

	precision	recall	f1-score	support
1	1.00	1.00	1.00	698
2	1.00	1.00	1.00	26868
3	1.00	1.00	1.00	777
accuracy			1.00	28343
macro avg	1.00	1.00	1.00	28343
weighted avg	1.00	1.00	1.00	28343



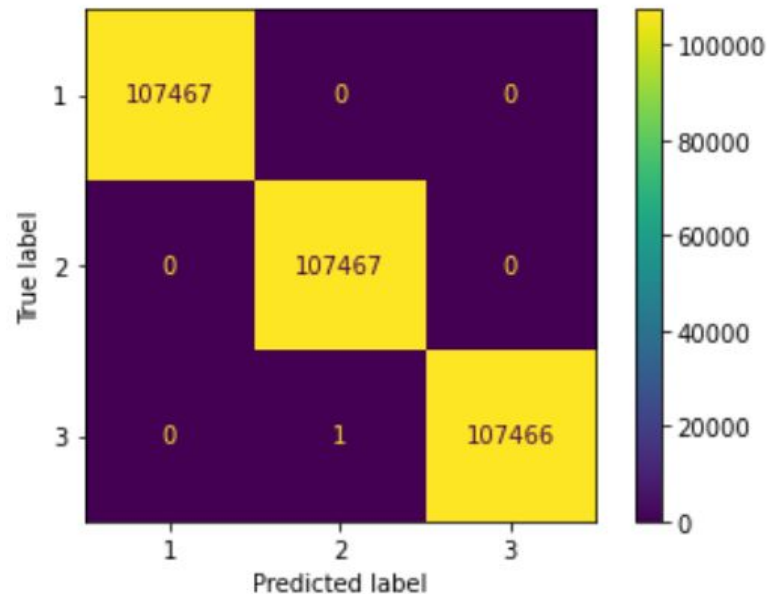
## Decision tree: Test data

	precision	recall	f1-score	support
1	1.00	1.00	1.00	698
2	1.00	1.00	1.00	26868
3	1.00	1.00	1.00	777
accuracy			1.00	28343
macro avg	1.00	1.00	1.00	28343
weighted avg	1.00	1.00	1.00	28343



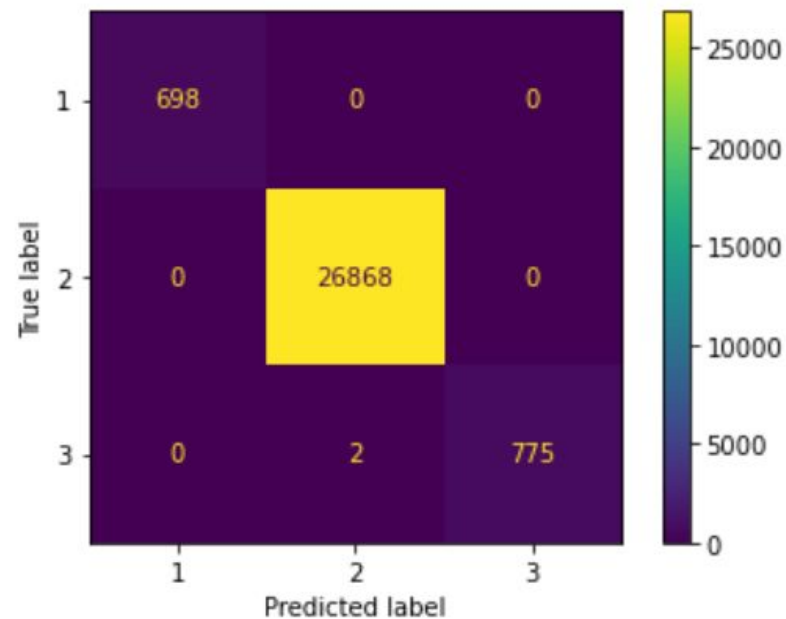
## Random forest: Train data

	precision	recall	f1-score	support
1	1.00	1.00	1.00	107467
2	1.00	1.00	1.00	107467
3	1.00	1.00	1.00	107467
accuracy			1.00	322401
macro avg	1.00	1.00	1.00	322401
weighted avg	1.00	1.00	1.00	322401



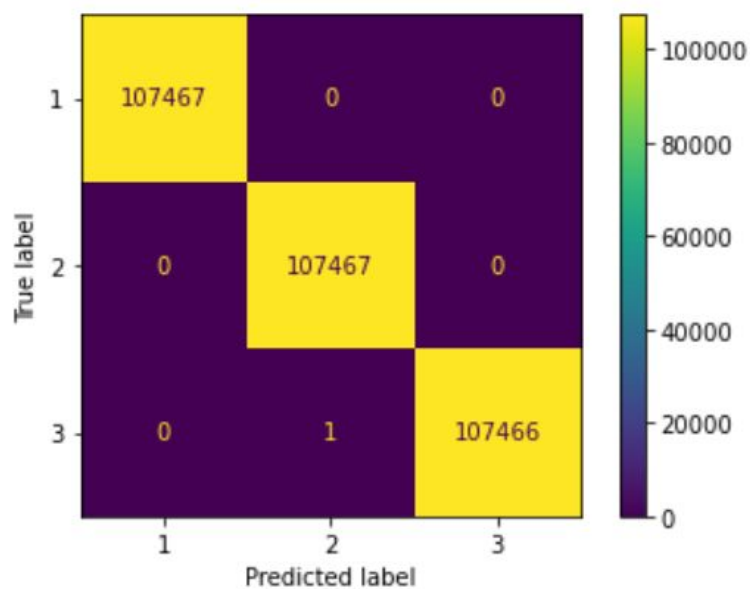
## Random forest: Test data

	precision	recall	f1-score	support
1	1.00	1.00	1.00	698
2	1.00	1.00	1.00	26868
3	1.00	1.00	1.00	777
accuracy			1.00	28343
macro avg	1.00	1.00	1.00	28343
weighted avg	1.00	1.00	1.00	28343



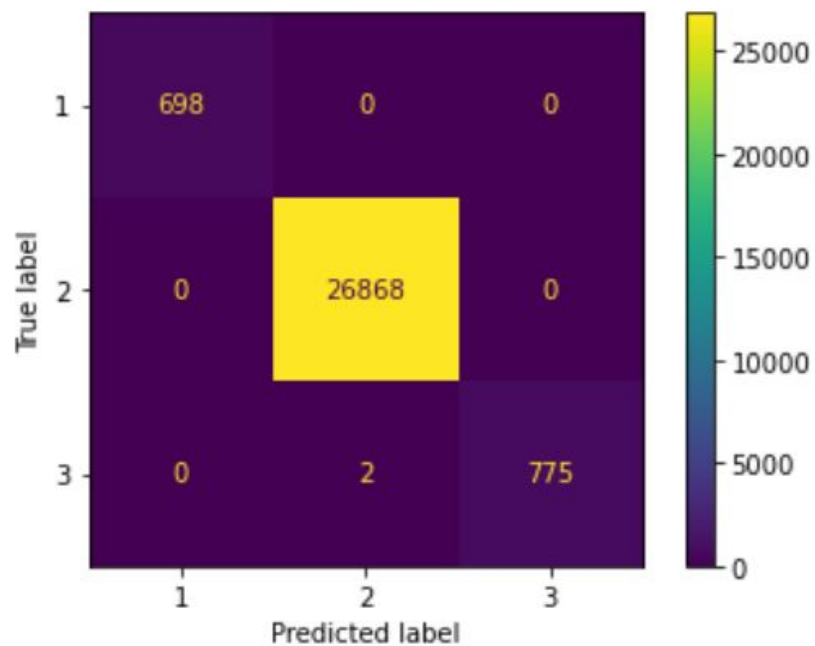
## XGBoost: train data

	precision	recall	f1-score	support
1	1.00	1.00	1.00	107467
2	1.00	1.00	1.00	107467
3	1.00	1.00	1.00	107467
accuracy			1.00	322401
macro avg	1.00	1.00	1.00	322401
weighted avg	1.00	1.00	1.00	322401



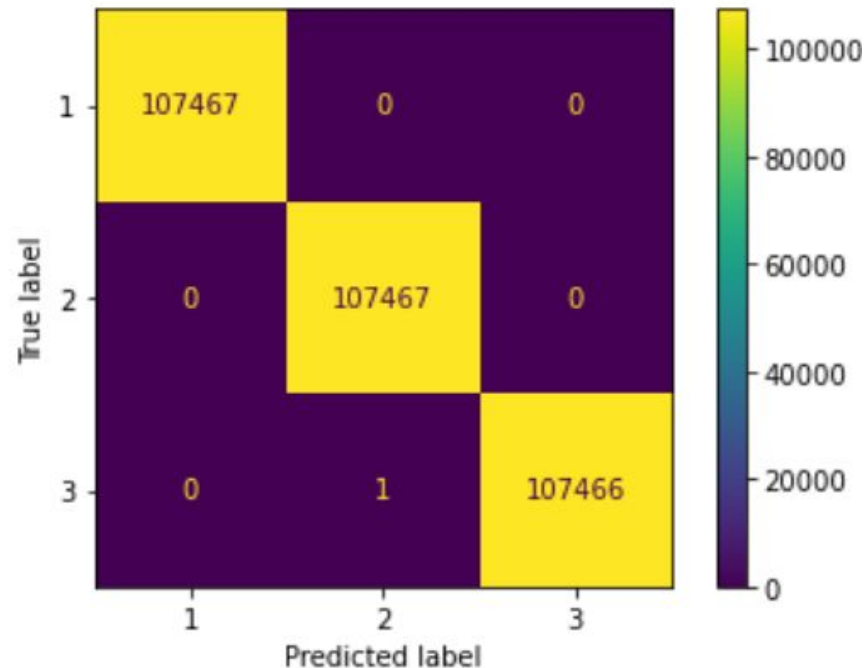
## XGBoost: Test data

	precision	recall	f1-score	support
1	1.00	1.00	1.00	698
2	1.00	1.00	1.00	26868
3	1.00	1.00	1.00	777
accuracy			1.00	28343
macro avg	1.00	1.00	1.00	28343
weighted avg	1.00	1.00	1.00	28343



## KNN: Train data

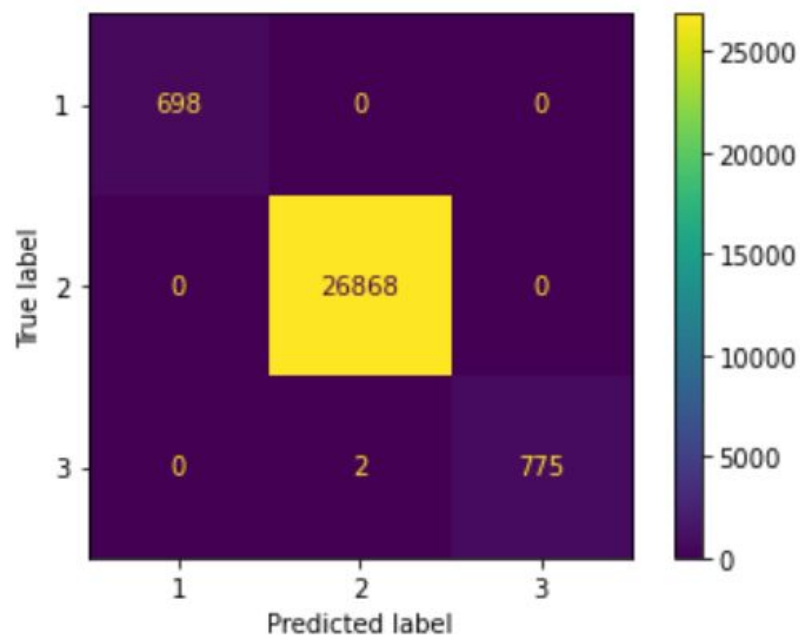
	precision	recall	f1-score	support
1	1.00	1.00	1.00	107467
2	1.00	1.00	1.00	107467
3	1.00	1.00	1.00	107467
accuracy			1.00	322401
macro avg	1.00	1.00	1.00	322401
weighted avg	1.00	1.00	1.00	322401





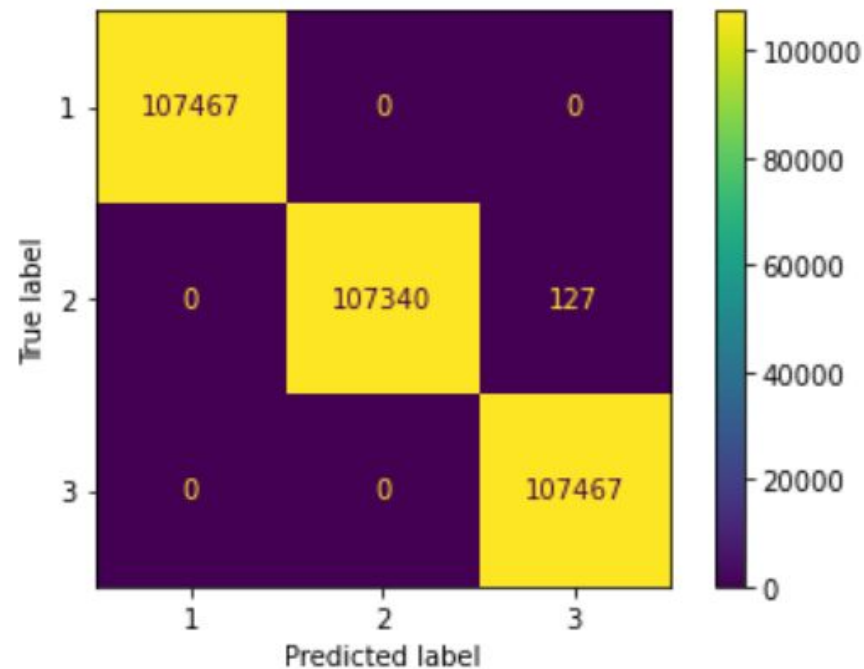
## KNN: Test data

	precision	recall	f1-score	support
1	1.00	1.00	1.00	698
2	1.00	1.00	1.00	26868
3	1.00	1.00	1.00	777
accuracy			1.00	28343
macro avg	1.00	1.00	1.00	28343
weighted avg	1.00	1.00	1.00	28343



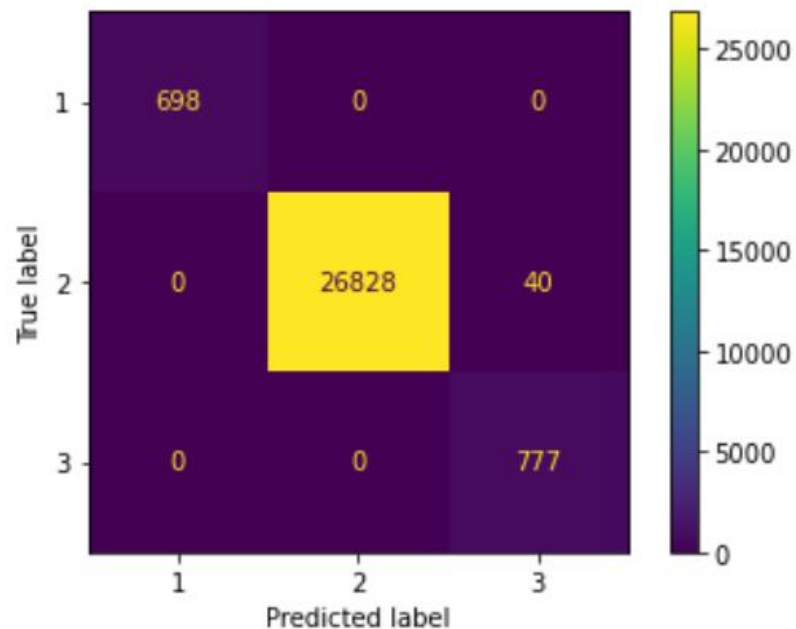
## SVM: Train data

	precision	recall	f1-score	support
1	1.00	1.00	1.00	107467
2	1.00	1.00	1.00	107467
3	1.00	1.00	1.00	107467
accuracy			1.00	322401
macro avg	1.00	1.00	1.00	322401
weighted avg	1.00	1.00	1.00	322401



## SVM: Test data

	precision	recall	f1-score	support
1	1.00	1.00	1.00	698
2	1.00	1.00	1.00	26868
3	0.95	1.00	0.97	777
accuracy			1.00	28343
macro avg	0.98	1.00	0.99	28343
weighted avg	1.00	1.00	1.00	28343



## Summary of results for all the models:

A weighted average of all the precision, recall , F1 score and accuracy is reported for each of the models mentioned.

Models	Dataset type	Weighted average of scores Train data				Weigthd average of scores Test data				Remarks
	Imbalanced dataset	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score	Accuracy	
Decision tree baselin		1	1	1	1	1	1	1	1	included all features into the model.
Decision tree		1	1	1	1	1	1	1	1	included only the features : 'urgency', 'priority', 'number', 'opened_by'
Random forest		1	1	1	1	1	1	1	1	
Xgboost		1	1	1	1	1	1	1	1	
Decision tree	1	1	1	1	1	1	1	1		
Random forest	1	1	1	1	1	1	1	1		
Xgboost	1	1	1	1	1	1	1	1		
Adaboost	0.83	0.67	0.56	0.67	0.95	0.97	0.96	0.97		
Catboost	1	1	1	1	1	1	1	1		
LGBM	1	1	1	1	1	1	1	1		
KNN	1	1	1	1	1	1	1	1		
SVC	1	1	1	1	1	1	1	1		
GNB	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99		
Logistic regression	1	1	1	1	1	1	1	1		

# Final model and results

## Model - Random Forest

**Data set details:**  
 features - categorical and numeric  
 target - 'Impact'

**Data Partition details:**  
 Train - 80%  
 Test - 20%

**Algorithm details and configuration:**  
 Data preprocessing  
 Feature engineering + selection  
 Random forest - chosen for stability and performance  
 No hyper parameters added

	Features selected	Impact Class	Precision	recall	f1-score	accuracy
Train	priority', 'number', 'urgency', 'opened_by'	1-High	1	1	1	1
		2-Medium	1	1	1	
		3-Low	1	1	1	
Test		1-High	1	1	1	1
		2-Medium	1	1	1	
		3-Low	1	1	1	

## **Possible Reasons for high accuracy (1 in all case)**

**Urgency:** Depending on the nature of the issue and after identifying the nature of impact, a decision is taken on how soon the issue is to be fixed. Impact severity is directly proportional to urgency.

**Priority:** Priority is fixed depending on the business service. Incident priority can be upgraded or downgraded based on the severity of the incident

**Impact:** It is decided based in terms of damage caused to business, type of customers impacted based on location and varies from business to business.

***All three are directly related to each other.***

**Note:** All possible ways for information leakage was checked and it was found that the same did not happen.


# Model Predictions

Model: Random forest

Transaction data  
fed to the model

Actual  
**impact**  
value

Predicted **impact**  
for each  
transaction

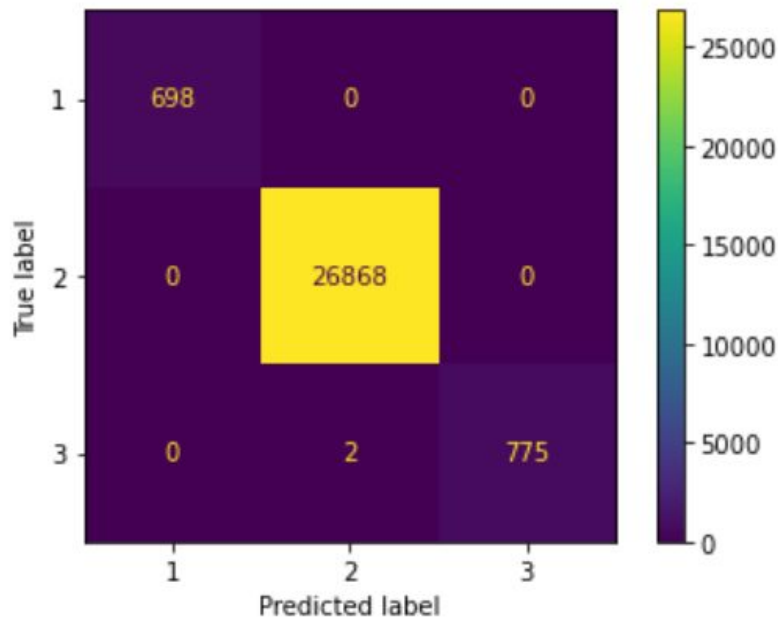


number	priority	urgency	opened_by	impact	impact_pred
29727	3	2	24	2	2
1287	3	2	468	2	2
15348	3	2	17	2	2
22969	2	2	305	1	1
30514	2	2	58	1	1
22377	2	2	24	1	1
33660	4	3	131	3	3
26073	4	3	20	3	3
2285	4	3	533	3	3

# Model Results

## Model: Random forest

	precision	recall	f1-score	support
1	1.00	1.00	1.00	698
2	1.00	1.00	1.00	26868
3	1.00	1.00	1.00	777
accuracy			1.00	28343
macro avg	1.00	1.00	1.00	28343
weighted avg	1.00	1.00	1.00	28343



- 698 correct predictions for 'High impact'.
- 26868 correct predictions for 'Moderate impact'.
- 775 correct predictions and 2 incorrect predictions for 'Low impact'



# **Model Deployment using Streamlit**

# Incident Impact Prediction - Web App

×

Input parameters

urgency

priority

number

opened by

Share

☆

☰

Incident Impact Predictor

Input parameters

	urgency	priority	number	opened_by
0				

Predict

Created by:-

Nithish , Sourajit

Rahul , Dhanya

Machindra , Vikas

Manage app

Link: [https://share.streamlit.io/niti42/incident\\_impact\\_v1.1/main/incident\\_pred\\_app\\_v1.1.py](https://share.streamlit.io/niti42/incident_impact_v1.1/main/incident_pred_app_v1.1.py)

# Incident Impact Prediction

## with sample data

✕

### Input parameters

urgency

2 - Medium

priority

2 - High

number

INC0000062

opened by

Opened by 180

Share ☆ ☰

## Incident Impact Predictor

### Input parameters

	urgency	priority	number	opened_by
0	2 - Medium	2 - High	INC0000062	Opened by 180

Predict

### Prediction

High impact

### Created by:-

Nithish , Sourajit

Rahul , Dhanya

Machindra , Vikas

< Manage app

# Challenges faced?

- Dominance of categorical features.
- Ambiguity in representation of null values.
- Date time features recorded using object datatype and had few null values.
- poor correlation between features and target except for a few features.
- imbalanced dataset.

# How did you overcome?

- Categorical features encoded with appropriate encoding techniques.
- All ambiguous cases considered as null values after multiple discussions with team, mentor and some analysis.
- Date time features converted to correct datatypes and appropriate features extracted.
- Poor correlations - Used feature selection techniques to eliminate unwanted features and consulted with a subject expert to verify if the finalized features were really important for predictions.
- Imbalanced dataset - Used SMOTE for balancing the dataset but performance was same with/without SMOTE.

**Thank you**