Data Science PROJECT

Client: ABC Tech | Category: ITSM - ML Project

Business Case:

ABC Tech is a mid-size organization operation in IT-enabled business segment over a decade. On an average ABC Tech receives 22-25k IT incidents/tickets, which were handled to best practice ITIL framework with incident management, problem management, change management and configuration management processes. These ITIL practices attained matured process level and a recent audit confirmed that further improvement initiatives may not yield return of investment.

ABC Tech management is looking for ways to improve the incident management process as recent customer survey results shows that incident management is rated as poor.

<u>Machine Learning as way to improve ITSM processes:</u>

ABC Tech management recently attended Machine Learning conference on ML for ITSM. Machine learning looks prospective to improve ITSM processes through prediction and automation. They came up with 4 key areas, where ML can help ITSM process in ABC Tech.

1. Predicting High Priority Tickets: To predict priority 1 & 2 tickets, so that they can take preventive measures or fix the problem before it surfaces.

2. Forecast the incident volume in different fields, quarterly and annual. So that they can be better prepared with resources and technology planning.

1. Predicting High Priority Tickets: To predict priority 1 & 2 tickets, so that they can take preventive measures or fix the problem before it surfaces.

- ## Assumptions:

  Identify the dataset as supervised categorical and numerical dataset.

  - ### Existing dataset:

    The following data is the original dataset, having **46606** rows (records) with some missing values and **25** columns (feature variables).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 46606 entries, 0 to 46605
Data columns (total 25 columns):
CI_Name                          46606 non-null object
CI_Cat                           46606 non-null object
CI_Subcat                        46606 non-null object
WBS                              46606 non-null object
Incident_ID                      46606 non-null object
Status                           46606 non-null object
Impact                           46606 non-null object
Urgency                          46606 non-null object
Priority                         46606 non-null object
number_cnt                       46606 non-null object
Category                         46606 non-null object
KB_number                        46606 non-null object
Alert_Status                     46606 non-null object
No_of_Reassignments              46606 non-null object
Open_Time                        46606 non-null object
Reopen_Time                      46606 non-null object
Resolved_Time                    46606 non-null object
Close_Time                       46606 non-null object
Handle_Time_hrs                  46606 non-null object
Closure_Code                     46606 non-null object
No_of_Related_Interactions       46606 non-null object
Related_Interaction              46606 non-null object
No_of_Related_Incidents          46606 non-null object
No_of_Related_Changes            46606 non-null object
Related_Change                   46606 non-null object
dtypes: object(25)
memory usage: 8.9+ MB
```

❖ Refined dataset:

- As some of the columns are not useful for Analysis and Modelling, we will remove them.
  Dropped columns are – 'Status','number_cnt','Alert_Status','Open_Time','Resolved_Time', 'Close_Time', 'Handle_Time_hrs', 'Related_Interaction', 'Related_Change'.
- As **'No_of_Related_Changes**' and **'No_of_Related_Incidents'** columns has more number of missing values and also not useful for prediction, we can drop the column.
- Adding a new column based on **'Reopen_Time'** as **'ReOpen_flag'** for the incidents which has been Re-Opened ➔ (1) or Not Re-Opened ➔ (0).
- As per the business problem, we need to predict High priority tickets i.e., Priority 1 and Priority 2.
  So we filter the feature variable **'Priority'** based on condition and add a new feature as **'New_Priority'** ➔ Target variable (y)
  'New_Priority' will contain 'High_Priority' and 'Low_Priority'.
  The Priority which is 1 and 2 are High_Priority.
  The Priority which is 3, 4 and 5 are Low_Priority.

➢ Top Features influencing the prediction of High priority tickets:

After Data cleaning, Exploration data analysis, Correlation and based on some domain knowledge, we came up with some features which will impact the High priority prediction.

Features	Description
CI_Name	Name of CI. CI is basically a service being used by the organizational user, it could be an application, infrastructure service etc.
WBS	WBSE Could be the department or equivalently a charge code for using the CI service.
No_of_Reassignments	It's the no. of Hands offs between Resolution Agents before the issue gets resolved. Generally High priority tickets should have less hand-offs to other agent. Re-assignments happen in following scenarios: 1) When the Agent at the time of creation assigns to wrong resolution team 2) When the issue require resolutions from multiple teams (like software, infra teams) and there is no one single owner that could be established.

	3) Sometimes issue is reassigned due to cursory or inadequate investigation.
Closure_Code	It could be a code to capture the cause of incident. This information is updated by the ticket resolution agent. Since at the time of capture, the ticket creation agent will not know about the information, it cannot be a predictor. However if we assume that the Ticket Creation agent can see the list of related incidents at the time of creating a new ticket, the closure code can be of some value to assess the priority.
No_of_Related_Interactions	Count of number of interactions with customer before the ticket was resolved. In general sense, if the priority of the tickets is higher, the SLA will be stringent and the Resolution agent may try to have many calls to get as much information required for reproduction of issue etc. to resolve it. Hence this could predict the priority
ReOpen_flag	Time Stamp when the ticket was reopened. Typically if the resolution agent closes/resolves the tickets without satisfaction of the user or the issue resurfaces, then the user re-opens the ticket and this time may be by increasing the priority, (ITIL doesn't recommend such behaviors). Reopen Time may not help but can be converted into binary yes/no to have the impact of Reopening on Priority

- **Machine Learning algorithms:**

Algorithm	**Accuracy**	**Recall**	**Precision**
Logistic Regression	76	73	76
K-Nearest Neighbor	83	79	86
SVM	83	80	84
MLP Classifier (ANN)	79	60	93
Decision Tree	83	67	96
Random Forest	86	79	91
XG Boost	87	84	89

- **Tools used:**

Jupyter - Python 3.7, MySQL, Excel, Machine learning algorithms.

➢ Results:

Algorithm XG boost classifier has given a good Accuracy with 87% and Recall rate with 84%, which is more efficient for Model prediction.

➢ Summary:

Our goal is to predict High priority tickets i.e., Priority 1 and 2 tickets. But the data here is more skewed towards low priority tickets and more number of data entries are among low priority tickets. There is an imbalance nature in the dataset. So here we used under-sampling technique to equalize the dataset among high priority and low priority tickets.

Before Under-sampling:

New priority	**Total count**
Low Priority	35634
High Priority	685

After Under-sampling:

New priority	**Total count**
Low Priority	685
High Priority	685