# Potential CFD Use Case

## Problem Statement

To predict the propensity of an internally found defect becoming a Customer found defect.

## Datasets

* We use three types of data for the model
  + Defect Data
  + Defect Text Data
  + Defect Lifecycle Data
* Below tables give details about each of these data sets

### Defect Data

|  |  |  |
| --- | --- | --- |
| Data Field | Data Field | Data Field |
| IDENTIFIER | DE\_MANAGER\_USERID | SEVERITY\_CODE |
| LIFECYCLE\_STATE\_CODE | FOUND\_PHASE\_CODE | REGRESSION\_BUG\_FLAG\_OLD |
| PROJECT | PRODUCT | COMPONENT |
| ENGINEER | SUBMITTER\_ID | PRIORITY\_DESC |
| PRIORITY\_CODE | FEATURE | SA\_ATTACHMENT\_TYPE |
| CR\_ATTACHMENT\_TYPE | UT\_ATTACHMENT\_TYPE | SA\_ATTACHMENT\_INDIC |
| CR\_ATTACHMENT\_INDIC | UT\_ATTACHMENT\_INDIC | CATEGORY |
| IMPACT | BADCODEFIXID | ORIGIN |
| ORIGINAL\_FOUND | OPENED\_ON | IS\_CUSTOMER\_VISIBLE |
| DUPLICATE\_OF | NOT\_CUSTOMER\_VISIBLE\_REASON | VERSION\_TEXT |
| TICKETS\_COUNT | INCOMING\_INDIC | BACKLOG\_INDIC |
| DISPOSED\_INDIC | SEV1\_INDIC | SEV2\_INDIC |
| SEV3\_INDIC | SEV4\_INDIC | SEV5\_INDIC |
| SEV6\_INDIC | TS\_INDIC | SS\_INDIC |
| OIB\_INDIC | STATE\_ASSIGN\_INDIC | STATE\_CLOSE\_INDIC |
| STATE\_DUPLICATE\_INDIC | STATE\_FORWARD\_INDIC | STATE\_HELD\_INDIC |
| STATE\_INFO\_INDIC | STATE\_JUNK\_INDIC | STATE\_MORE\_INDIC |
| STATE\_NEW\_INDIC | STATE\_OPEN\_INDIC | STATE\_POSTPONE\_INDIC |
| STATE\_QA\_INDIC | STATE\_RESOLVE\_INDIC | STATE\_SUBMIT\_INDIC |
| STATE\_UNREP\_INDIC | STATE\_VERIFY\_INDIC | STATE\_WAIT\_INDIC |
| PRIORITY1\_INDIC | PRIORITY2\_INDIC | PRIORITY3\_INDIC |
| PRIORITY4\_INDIC | PRIORITY5\_INDIC | PRIORITY6\_INDIC |
| PRIORITY7\_INDIC | PRIORITY8\_INDIC | PRIORITY9\_INDIC |
| CFD\_INDIC | IFD\_INDIC | CFR\_INDIC |
| SS\_EVAL\_INDIC | SP12\_INDIC | MTTR |
| S3P12\_INDIC | RD\_INDIC | INVALID\_SS\_INDIC |
| POTENTIAL\_SS\_INDIC | S12RD\_INDIC | S12CFD\_INDIC |
| S123RD\_INDIC | MISSING\_SS\_EVAL\_INDIC | S123\_INDIC |
| S12\_INDIC | CREATED\_IN\_LAST\_14D\_INDIC | DEV\_ESCAPE |
| TEST\_ESCAPE | RNE\_INDIC | CREATED\_BY |
| UNREPRODUCIBLE | BUG\_ORIGIN | DETAILED\_ACTIVITY |
| DEV\_ESCAPE\_ACTIVITY | FOUND\_DURING | ORIGINAL\_FOUND\_DURING |
| RELEASED\_CODE | RELEASED\_CODE\_ID | TEST\_EDP\_ACTIVITY |
| TEST\_EDP\_COMMENTS | TEST\_EDP\_PHASE | ACTIVITY\_WHEN\_FOUND |
| RESOLVER\_ANALYSIS\_INDIC | SUBMITTER\_ANALYSIS\_INDIC | EDP\_ANALYSIS\_INDIC |
| RETI\_ANALYSIS\_INDIC | DESIGN\_REVIEW\_ESCAPE\_INDIC | STATIC\_ANALYSIS\_ESCAPE\_INDIC |
| FUNC\_TEST\_ESCAPE\_INDIC | SELECT\_REG\_ESCAPE\_INDIC | CODE\_REVIEW\_ESCAPE\_INDIC |
| UNIT\_TEST\_ESCAPE\_INDIC | DEV\_ESCAPE\_INDIC | FEATURE\_TEST\_ESCAPE\_INDIC |
| REG\_TEST\_ESCAPE\_INDIC | SYSTEM\_TEST\_ESCAPE\_INDIC | SOLUTION\_TEST\_ESCAPE\_INDIC |
| INT\_TEST\_ESCAPE\_INDIC | GO\_TEST\_ESCAPE\_INDIC | IFD\_CFD\_INDIC |
| COMPLETE\_ESCAPE\_INDIC | SR\_CNT | PSIRT\_INDIC |
| BADCODEFIX | BADCODEFLAG | SR\_NUMBERS |
| RISK\_SCORE | NORMALIZED\_RISK\_SCORE | RISK\_LEVEL |
| RISK\_OWNER | SIR | PSIRT\_FLAG |
| URC\_DISPOSED\_INDIC | CLOSED\_DISPOSED\_INDIC | REGRESSION\_BUG\_FLAG |
|  |  |  |

### Defect Text Data

|  |  |  |
| --- | --- | --- |
| Data Field | Data Field | Data Field |
| HEADLINE | ATTRIBUTE | SS\_EVAL\_TEXT |
| RETI\_ANALYSIS | EDP\_ANALYSIS | RESOLVER\_ANALYSIS |
| SUBMITTER\_ANALYSIS | DESCRIPTION |  |
|  |  |  |

### Defect Lifecycle Data

* Date when the defect got converted from an internally found to customer found

## Exploratory Data Analysis(EDA)

### Data Preprocessing

#### Defect Data Preprocessing

* The preprocessing of the data includes imputation of missing values in attributes, removing the outliers and handling the categorical attributes.
* During this phase, we removed attributes that accounted to more than 70% missing data, date attributes and redundant attributes.
* Below is the list of few attributes that we excluded for the analysis.
* Redundant attributes like PRIORITY1\_INDIC and others are removed because we already considered PRIORITY\_CODE
* Attributes like FOUND\_DURING, CFD\_INDIC and IFD\_INDIC are removed as they bias the model predictions

|  |  |  |
| --- | --- | --- |
| Data Field | Data Field | Data Field |
| PRIORITY\_DESC | RELEASE\_NOTE | SA\_ATTACHMENT\_TYPE |
| CR\_ATTACHMENT\_TYPE | UT\_ATTACHMENT\_TYPE | CLOSED\_ON |
| PRIORITY1\_INDIC | PRIORITY2\_INDIC | PRIORITY3\_INDIC |
| PRIORITY4\_INDIC | PRIORITY5\_INDIC | PRIORITY6\_INDIC |
| PRIORITY7\_INDIC | PRIORITY8\_INDIC | PRIORITY9\_INDIC |
| CFD\_INDIC | IFD\_INDIC | CREATED\_IN\_LAST\_14D\_INDIC |
| SS\_EVAL\_TEXT | CREATED\_ON | UPDATED\_ON |
| HELD\_ON | INFO\_REQ\_ON | JUNKED\_ON |
| NEW\_ON | POSTPONED\_ON | VERIFIED\_ON |
| WAITING\_ON | MORE\_ON | LAST\_MOD\_ON |
| FIRST\_CLOSED\_ON | FOUND\_DURING | ORIGINAL\_FOUND\_DURING |
| BADCODEFIX | TEST\_EDP\_COMMENTS | DEV\_ESCAPE\_ACTIVITY |

Refer to EDA PDF or [http://localhost:8888/notebooks/CSC\_ena.ipynb#](http://localhost:8888/notebooks/CSC_ena.ipynb)

### Data Sampling

Whole data is split into Train and Test sets based on an input. 3 years of data is split to 2 years of training and 1 year of testing data. Training data is used to train the model and testing is used to cross-validate model prediction results.

By default, training set is 2016 and 2017, while testing set is 2018.

### Feature Selection

* This is the key phase for building high performance models. In this phase we found the most important predictor attributes that explain major part of variance of the response variable
* This feature selection is done using Xtreme-Gradient boosting algorithm. The model is hyper tuned on the parameters and a grid search is performed.
* Out of 165+ attributes the model was able to filter out approximately 30+ features for each model.
* Below is the list of a few features that were considered important by the model.

|  |  |  |
| --- | --- | --- |
| Data Field | Data Field | Data Field |
| DE\_MANAGER\_USERID | AGE | ENGINEER |
| SUBMITTER\_ID | FEATURE | IMPACT |
| LIFECYCLE\_STATE\_CODE | COMPONENT | ORIGIN |
| SEVERITY\_CODE | PRIORITY\_CODE | TICKETS\_COUNT |
| INCOMING\_INDIC | BACKLOG\_INDIC | UNIT\_TEST\_ESCAPE\_INDIC |

## Model Architecture

### Classification of IFDs

* For the prediction of propensity of an IFD becoming CFD, we use 3 models.
* The first model is built on the categorical and continuous attributes. Approximately 30+ attributes that we got from the feature selection are used.
* Xtreme-Gradient Boosting mechanism is used for this model. The model goes through grid search over the parameters for hyper tuning.
* This boosting model outputs a probability with which it says a bug can become a CFD.
* The second model is built on textual data.
* Fields like ENCL\_DESCTIPTION, HEADLINE and ATTRIBUTE are considered. The textual data also includes log files, email attachments or chat conversations.
* The text classification is built using Convolutional Neural Networks and LSTM.
* The textual data is encoded into vectors. These vectors are embedded into matrix, which are then fed to the feed forward convolutional neural network.
* CNN is used to figure out the straight dependencies in the sequence of words in the sentences. This model outputs the invariant word features.
* The invariant features are passed to the Long Short Term Memory RNN. This model learns the sequence patterns by reducing the cross entropy loss.
* This model outputs the probability with which it can say if the bug becomes CFD.
* The final model is an ensemble method that uses both the probabilities from the classification models and makes a final prediction.
* The training set is bugs from January 2016 till July 2017, while the testing set is bugs from August 2017 till date. The training set accounted for 70% of the total data.

### Prediction of lead time

* After classifying whether a bug has the propensity to become a potential CFD, we compute the lead time within which the customer might face this issue.
* This lead time prediction is modelled using Deep Neural Networks model.
* This model architecture takes the features as input and predicts the time (in days), from the day the bug was submitted.