OVERVIEW OF THE FINAL PROCESS

BUSINESS PROBLEM

Classification of tickets and assignment to their appropriate resolver groups based on available ticket data and features

OUR APPROACH

Division of data into two groups, one for groups with low number of records and the other for groups with high number of records

SELECTED STRATEGY ELEMENTS

Features of Data:

- Ticket descriptions are Multilanguage
- Data is highly imbalanced
- Data has text as well as numeric features

Our preprocessing steps:

- Extract numeric features
- Cleaning data, encodings using FTFY
- Translating records, Augmenting the data

Translation used:

- G-translate loops with backout
- Translation runs before and after data cleaning

Augmentation used:

- Random augmentation of upto 10 words in a record
- Random word dropping at ~15%

Algorithms used:

- Logistic Regression, SVM, XGB
- Random Forest, LightGBM, CatBoost
- GRU, LSTM, Bidirectional LST

Techniques used:

- Skip connections in GRU
- Joining LSTM text and Numeric models
- Running data through multiple models to maximize precision and recall

SOLUTION ROADMAP: STEP BY STEP

	ACTION	FINDING	DETERMINATION
Step 1	• Perform EDA on data to discover data challenges	• Imbalanced target group distribution, garbage words & multilanguage source data	• Need for Data, cleaning translation & augmentation
Step 2	• Perform Data cleaning on the dataset to remove fluff words and garbage	• Data encoded in UTF-8	• Requires conversion into corresponding language codes through FTFY
Step 3	• Initiate Language Translation through google translate libraries	• Partial translations on first attempt	• Greater translation requires removal of additional markers such as <Chinese text>,34323 → detected as English

SOLUTION ROADMAP: STEP BY STEP

	ACTION	FINDING	DETERMINATION
Step 4	• Augment data to create a larger dataset	• Standard augmentation with synonyms fails to capture latent relationships	• Inclusion of drop words augmentation
Step 5	• Demarcate datasets into rule-based and AI-based and perform vectorization	• Increasing features increases accuracy for TFIDF, additionally, some feature engineering also helps	• Improved feature extraction required, and further combination of models needed
Step 6	• Perform machine learning and deep learning on data	• Accuracy lacking in some target groups	• Improve feature extraction, translation and data cleaning

MODEL EVALUATION

FINAL MODEL

- ✓ Random Forest

BEST PARAMETERS

- ✓ 'criterion': 'entropy'
- ✓ 'max_depth': None
- ✓ 'max_features': 'log2'
- ✓ 'n_estimators': 50

PROMINENT PARAMETERS

- ✓ Increasing n_estimators directly reduces overfitting of the model and increases test accuracy

Objective

- ➢ High overall accuracy on test set
- ➢ High individual accuracy on target classes
- ➢ Strong balance between precision and recall

Evaluation of Success

- ➢ Performance on test data
- ➢ Performance on validation data
- ➢ Performance on re-augmented test set
- ➢ Individual class-wise accuracy/precision/recall above threshold

INTERIM PERFORMANCE OF MODELS

Machine Learning

Logistic Regression

- Fast training and scales well

Performance

- 84% (Single Model)
- 90% (Tuned Model)

XGB

- Known for accuracy and stability of results

Performance

- 84% (Single Model)
- 80% (Tuned Model)

SVC

- Works well with text data

Performance

- 84% (Single Model)

Deep Learning

GRU

- Fast training and scales well

Performance

- 91% (Base Model)

LSTM

- Offers increased accuracy over GRU
- Marginally slower

Performance

- 89.3% (Base Model)

Bidirectional LSTM

- Has not shown significant gain over LSTM
- Much slower

Performance

- 85% (Base Model)

We were unable to hyperparameterize XGB and SVC due to the significant time they were consuming. For deep learning, we will start the tuning in the next phase. The low performance of LSTM and Bidirectional LSTM is solely due to the need for greater training time.

OUR INTERIM PLAN ON FUTURE OPTIMIZATIONS

Improving our outcomes can be done with 4 broad themes

OPTIMIZATION THEME	OPTIMIZATION TOPICS
Improving the Data	• Increasing % of translation. Translating by sentence approaches • Adding more variety in augmentation through Language generative ML
Hyperparameter Tuning	• Experimentation with different ranges of Hyperparameters with Random Search • Breaking up the training into phases and then selecting the right parameters for each phase
Feature Engineering	• Discovering new approaches of feature engineering such as topic modelling, summary generation etc. • Adding statistical columns to support ML
Latest Models	• Applying latest models and capabilities for Deep Learning and Machine learning such as Attention mechanisms

OUR FINAL MODELS

Classification Model	Training Accuracy	Test Accuracy
Logistic Regression	90.74	89.62
Logistic Regression using Random Search	94.32	93.57
Logistic Grid Classifier	94.41	93.64
XGBoost Classifier1	91.68	90.08
XGBoost Classifier2	95.78	94.42
XGBoost Classifier Grid Search	96.60	95.52
Support Vector Classification	94.03	93.45
Support Vector Classification (Grid Search)	96.98	96.91
Random Forest Classifier	98.41	97.78
Random Forest Classifier (Grid Search)	98.46	98.31
LightGBM	98.28	97.52
LightGBM (Grid Search)	98.35	97.62
CatBoost Classifier	89.58	88.10
CatBoost Classifier Grid	82.34	80.54
CatBoost Classifier Best	92.41	90.96
LSTM Merged Model	94.99	94.26
LSTM	94.89	94.42
GRU	95.34	94.85
Bi Directional LSTM	95.39	94.84
GRU Skip Connection	95.63	95.28
Random Forest Classifier Base Data	98.46	65.61
LightGBM Base Data	97.89	68.18
Logistic Regression Base Data	82.18	60.29

Models selected

Models on base data

COMPARISON TO BENCHMARK

WE DELIVERED OVER 7% IMPROVEMENT TO OUR 91% BENCHMARK IN THE INTERIM

1 IMPROVEMENT IN MODELS

Leveraging Skip Connections, Merged deep learning models, we were able to improve deep learning accuracy past 95%

2 IMPROVED FEATURIZATION

Compared to our interim efforts, we extracted multiple new features of value to the classification

3 IMPROVED DATA TREATMENT

We improved our augmentation outcomes and variety, increased the degree of translation to almost 100% and reduced the data loss from cleaning

4 IMPROVEMENT IN TUNING

We brought our hyperparameters closer to optimal ranges using coarse and fine gradations from the default values

Our final accuracy was above 98% using random forest models, this is a 30% improvement over the translated data without augmentation

IMPLICATIONS FOR BUSINESS

FASTER ASSIGNMENT TO GROUPS

The solution proceeds through the test group (19744 records) in around 10 minutes for 3 models. This means that assignment is being performed at 2000 records per minute. Compared to human assignment, this is several hundred times faster

ACCURACY OF ASSIGNMENT

The best model of the lot is random forests, which provides 98.31% accuracy. Combining this with other deep learning models, we can move towards nearly 99% accuracy which would be a significant improvement over human assignment

WE RECOMMEND

We recommend the below

1. Dissolution of groups which have less than 1% records from the total
2. Auto resolution of several tickets that relate to SID's or Hosts or batches that have failed. They can be automatically restarted based on detection

LEVEL OF CONFIDENCE

Based on the set of values we obtain from our best models; we can estimate the accuracy to be within 96.3% - 98.61%, 99% of the times.

LIMITATIONS OF THE MODEL

LIMITATIONS, REAL WORLD DEFICIENCIES, AND FUTURE ENHANCEMENTS

LIMITATION:

CHINESE LANGUAGE

Currently, the caller is a key differentiating factor in the tickets raised in Chinese language. However, as tickets increase, there could be some overlap between the users in different groups causing accuracy reduction.

Additionally, tickets raised automatically in groups 5,6,8,9 are challenging for the algorithm to classify

DEFICIENCY:

LIMITED DATA

Real world data can be vastly different from the data provided. In production scale ticketing systems, organizations can get hundreds of thousands of tickets per year. This means that we would not have captured the entire vocabulary for real time assignment, nor planned for its full impact on performance & future training

ENHANCEMENT:

CONTINUED LEARNING FROM CLOSED TICKETS

We need to configure the solution in a way that allows models to continuously be trained on new data as it arrives

The solution also needs to generate new patterns of language and vocabulary to create a deeper coverage of what a caller may raise in a ticket

CLOSING REFLECTIONS

	OBSERVATION	LEARNING FOR FUTURE
DATA	LOW ACCURACY IN 6% OF TARGET GROUPS DISCOVERED POST MODEL BUILDING	GROUP BY GROUP RECORD REVIEW IN AS PART OF INITIAL EDA
DATA	ACCURACY REDUCTION DUE TO DATA LOSS IN CLEANING DATA	DETERMINATION OF VALUE OF EVERY FEATURE/COLUMN PRIOR TO CLEANING
MODEL	LARGE DEVIATION FROM BASE HYPERPARAMETERS DURING TUNING CAUSING LOSS OF EFFORT	GRADUAL GRADATION OF HYPERPARAMETERS FROM DEFAULT SETTINGS
MODEL	A SINGLE MODEL DOES NOT SUFFICE FOR ALL INFORMATION	STACKING OR MERGING MODELS MAY IMPROVE ACCURACY