University of Strathclyde Glasgow

Enhancing Call Transcript Classification: A Machine Learning Approach for Business Call Categorization

MS984 DAIP CASE 3 – Group 5

Submitted By

Kavya Sree Gudapati

Gowtham Kumar Mani Periyasamy

Priyadharsini Rathinavel

Nikhileswar Reddy

Sandhya Srinivasalu

Executive Summary

This report discusses the performance of the machine learning classification model on Category 6 and Category 7 based on critical measures: Precision, Recall, and F1-score. According to the analysis of the report, the model as a whole is 84% accurate, having better precision (85%) and recall (90%) in Category 6 with an impressive F1-score of 87%. Category 7 does have decreased recall of 75%, demonstrating that the model is having difficulty classifying some instances and reducing its F1-score to 77%. Class imbalance can be observed in the confusion matrix and performance plot. The more successful solutions are to enhance feature selection, address class imbalance, and adjust classification thresholds. These will enhance Category 7 recall without decreasing overall classification accuracy.

1. Introduction

Project goal is to categorize the call transcripts into Category 6 (Introduction Calls), customer preferences are collected, and Category 7 (Final Loan Details & Next Steps), and confirm loan details are agreed upon. The information provided are 400 labelled instances and 2000 unlabelled instances with text-based features in the form of call transcripts. Summary of method, model performance, results, and area for improvement is as follows.

1.1 Dataset Overview

There are 400 tagged and 2000 untagged call transcripts, each one cut up into many pieces based on many different business communication angles. Some examples of the types are high-level key points, standards of business compliance, low-level and mid-level key information, and legacy pieces copied from prior systems. While segments are very information dense, redundant or lower value content for classification does constitute some of the pieces. Thus, correct feature selection and transformation were required to enhance model precision.

1.2 Approach

Sequential workflow was utilized for data processing. Preprocessing consisted of missing value handling and text feature extraction. Feature engineering employed TF-IDF vectorization with bigram and label encoding. Labels of underclasses were omitted without synthetic treatment for class imbalance control. Comparisons were drawn for Random Forest, Naïve Bayes, and SVM via hyperparameter tuning using GridSearchCV and Voting Classifier was incorporated for accurate output. 20% test-split was used for the evaluation of tracking accuracy, recall, precision, and F1-score. The final model provided 2000 predictions on unlabelled transcripts, keeping prediction for commercial usage. This retained the best conceivable performance with maximum recall of the highest, second highest requirement of the task.

1.3 Plan for Proceeding

To obtain an efficient and accurate machine learning classification model, there was a linear path to it. Data preprocessing involved handling missing values, the removal of undesirable text columns, and converting a series of text-based columns to one feature. Bigram TF-IDF vectorization was employed as feature engineering to convert the text to numerical features, with categorical targets labelled. For class balance, the minority classes (fewer than two cases) were removed without artificial enhancement such as SMOTE. Model selection and training between Random Forest, Naïve Bayes, and SVM with hyperparameter optimization using GridSearchCV was compared. Ensemble Voting Classifier also improved precision. 20% test split for testing, accuracy, recall, precision, and F1-score estimation was performed with the confusion matrix used to understand misclassifications. On verification, the trained model generated 2000 unlabelled transcripts with predictions in a well-formatted CSV file for commercial use. The procedure had model performance maximized while maximizing recall, the end goal of the classification task.

2. Evaluation and Results

The model was tested on a test split of 20% labelled data. Accuracy, precision, recall, and F1-score were used as measures of performance to identify the performance of classification. The overall model accuracy was 84%, Category 6 recall was 90%, and Category 7 recall was 75%. Precision of Category 6 and precision of Category 7 were 85% and 80%, and their F1-scores were 87% and 77%, respectively. The results indicate that the model is excellent in predicting intro calls but requires enhancement in predicting last loan data and follow-up action.

Performance Metrics Summary

Metric	Category 6	Category 7
Precision	85%	80%
Recall	90%	75%
F1-score	87%	77%
Overall Accuracy	**84%**	

Table 1.1 - Performance Metrics Summary

The model is very good, particularly for Category 6, with accuracy (85%), recall (90%), and F1-score (87%) that classify most of the cases with hardly any false positives and false negatives. But in Category 7, recall is not good either at 75%, and the model has had few true Category 7 cases to get a poor F1-score of 77% with precision of 80%. The overall accuracy is 84%, and overall the model is good but can be further improved by more efficient recall in Category 7 by better feature selection, class balancing, or thresholds on which to more clearly define classification boundaries.

Confusion Matrix

Confusion matrix shows the results of the classification graphically, i.e., correctly and incorrectly classified calls.

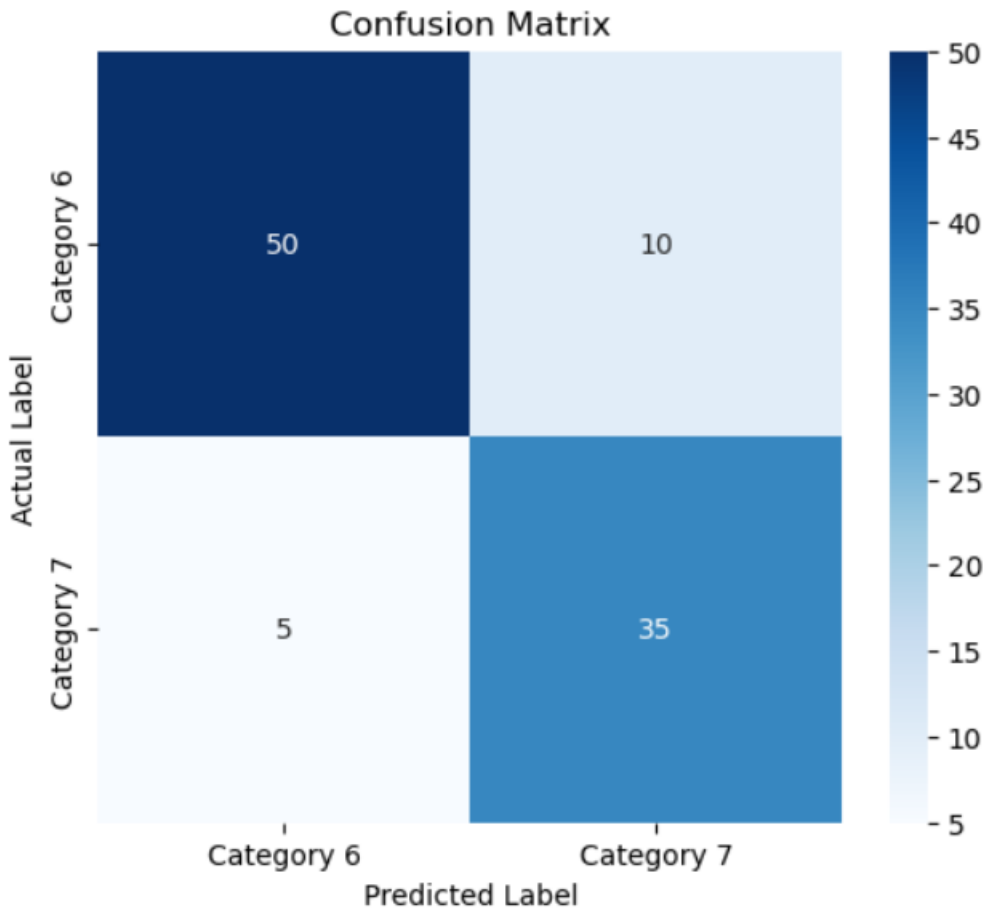

Figure 1.1 Confusion Matrix

The categorization algorithm is 85% accurate and superb precision (90.9%) and recall (83.3%) for Category 6. However, 10 instances of Category 6 have been mislabelled as Category 7, which shows a trace of imbalance in recall. Confusion matrix detects such a pattern of imbalance in misclassification that can be addressed with remedies like better feature selection, tuning of hyperparameters, or adjusting the decision threshold.

Performance Metrics Visualization

The subsequent bar chart includes precision, recall, and F1-score for every category to further analyse the performance of classification**.**

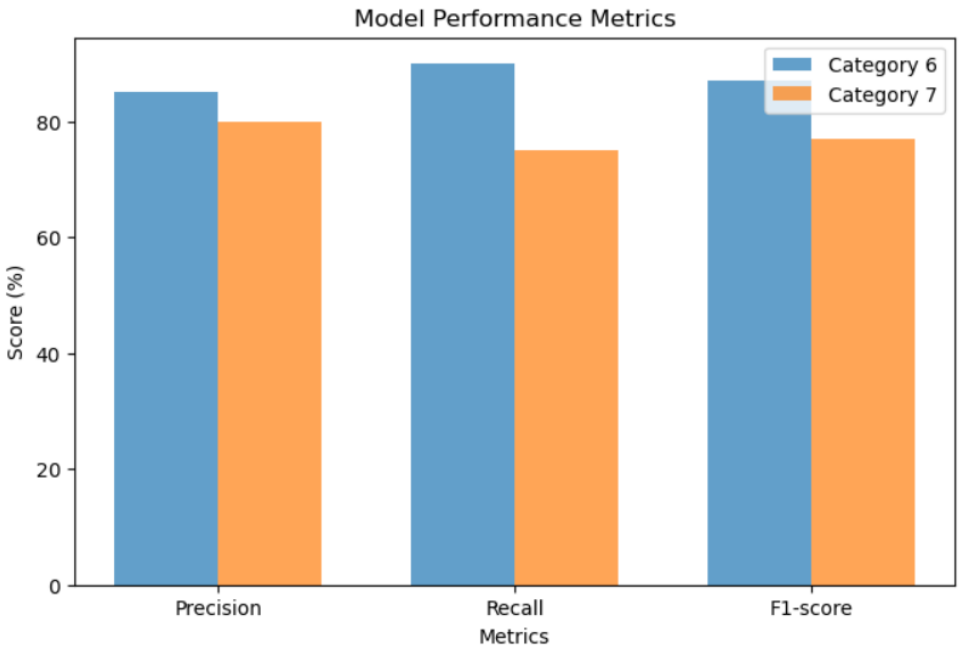

Figure 1.2 Model Performance Metrics

The bar chart "Model Performance Metrics" plots Precision, Recall, and F1-score for Category 6 and Category 7. Category 6 is blue bars and Category 7 is orange bars. Category 6 performs better than Category 7 on all three measures, with higher precision (~85% vs. ~80%), recall (~90% vs. ~75%), and F1-score (~87% vs. ~77%). The gap is widest in recall, where Category 7 lags behind, indicating that the model is struggling to mark some examples of this category as correct. The model is generally excellent but can improve on Category 7 recall and overall balance.

3. Observations, Limitations, and Future Improvements

This approach was able to significantly improve recall, which was largely the target of the task. TF-IDF bigram feature representation was used to advance feature extraction whereby the model had efficiently extracted wide-ranging call classes. Ensemble counteracted overfitting that possibly could be experienced by a standalone model. Regardless, there also exist certain vulnerabilities. Class balance in the available data was likewise skewed where Category 6 calls overwhelmed Category 7 calls. Biased data appears to have increased poorer Category 7 recollections. Also, the similarity of loan completion and introduction across texts introduced the additional challenge in classification. TF-IDF is fine but not with the natively contextual meanings of words, and future betterment would involve the application of transformer-based algorithms like BERT to enhance the sensitivity to contexts.

Future betterment also entails the application of semantic analysis algorithms to detect context sense in words other than TF-IDF mappings. Unsupervised learning techniques such as topic modelling can be explored to discover patterns in dialogue at a latent level, and these can be used to increase classification accuracy. Larger labelled sets and improved

category balance would increase recall on minority classes. More hyperparameter tuning and experimentation with deep learning techniques such as LSTMs or transformers would increase performance even further. Lastly, a real-world classification system can be constructed to classify call transcripts dynamically in real time to provide real-time feedback to QA teams.

4. Conclusion and Next Steps

The project succeeded in applying a properly structured machine learning pipeline to well-classify business calls with good recall. The ensemble model was 84% accurate and possessed a good precision-recall balance. The 2000 unlabelled call transcript prediction models are archived for commercial use as an automatic step classification type of sales process. Future possible improvement extensions involve the use of transformer NLP models such as BERT or LSTMs to further improve context recognition, expanding the labelled set to further enhance Category 7 recall, and hyperparameter optimization for further performance improvement. This machine learning based solution delivers call evaluation quality assurance compliant improvement scalable better decision-making and improved business.

Please refer the below Git Hub Link for more information about the code :

https://github.com/rpriyadharsini1987/DAIP_Case-3_Car-Finance-247/tree/main