DEPARTMENT OF

MANAGEMENT SCIENCE

Car Sales Call Classification: A Machine Learning Approach

by

202490978

202476189

202461546

202476957

MSc Data Analytics

MS984 Data Analytics in Practice

1. Introduction

1.1. Project Overview

This project focuses on classifying car sale transcripts obtained from phone calls. Currently at Car Finance 247, after being automatically forwarded the transcripts, an API interprets both incoming and outgoing calls according to responses to 200 pre-defined questions.

Despite this seemingly fluid system, its autonomous nature may not be optimal, especially within two specific call categories: Category 6 and Category 7 calls. Category 6 calls refer to introduction calls, where agents gather customer preferences and explain the intricacies of the loan application process. Conversely, category 7 calls are confirmation calls following completion of the loan agreement.

Quality assurance and compliance are something that cannot be ignored in modern business. Currently, the classification process obtained via API relies on heuristics, which, while maintaining an impressive 80% accuracy rate, has an extremely low recall of 2%. This means that when the system does recognise a call it is highly accurate at classifying it, however, it often fails to recognise Category 6 and 7 calls completely.

Missing these classifications can lead to gaps in compliance monitoring and hinder customer experience; thus, a new approach is required.

1.2. Objective and Rationale

The goal of this project is to create a machine learning model that eliminates this poor rate of recall while maintaining the rate of accuracy as much as possible. For a holistic approach, two models will be adopted in this project: a Supervised Machine Learning Model and a Semi-Supervised Machine Learning Model.

Due to the variability of sales calls, we have hypothesised that the semi-supervised approach will prove more effective. While the supervised model will provide a valuable benchmark, we believe that the semi-supervised model will have a greater capacity for generalisation and, therefore, reducing the likelihood of overlooking important calls.

2. Data Exploration and Pre-Processing

2.1. Dataset Overview

The provided data is divided into two datasets: one with 400 labelled transcripts with predefined categories and one with 2000 unlabelled transcripts. Our group were quick to identify that the former would be used to train the models and the latter to implement and enhance them.

2.2. Preprocessing Steps

Prior to analysis and the creation of any machine learning models, the data had to be scanned for missing data and cleaned if necessary. Through simple analysis, the data was confirmed to have no 'N/A' or missing responses throughout either dataset, so no imputation was required.

Following this, the text was cleaned throughout the dataset to facilitate analysis. A 'cleaner' function was defined to assist with this. This function implemented a 'for loop' which removed unnecessary brackets and quotation marks, along with – in the labelled dataset – classifying responses out with Category 6 or 7 as 'other'. Moreover, this function extracted key-value pairs from nested fields into separate columns and replaced categorical text outcomes such as 'TRUE' and 'FALSE' into numerical values, to facilitate analysis.

Following these steps, and a final missing values check, the cleaned data was exported into two new datasets; one for the unlabelled data and one for the labelled data.

3. Supervised Machine Learning Model

3.1. Model Selection

After a vigorous review of the cleaned datasets, our group decided that for the supervised model, a Random Forest Classifier was the most appropriate. There were various justifications for this. Not only was this a personal preference, but we knew that this type of model can handle a large number of features and is strongly resistant to overfitting. Therefore, implementation of a Random Forest model will allow us to handle the 147 features of the dataset, as well as avoiding the model being too stringently trained on the labelled dataset to the point where it cannot help us with the unlabelled data.

3.2. Training and Evaluation

The first step – prior to training – was to ensure a balanced approach. This was done by splitting the labelled dataset 80/20 into training and testing subsets respectively, this ensured that all three classes ('Category 6', 'Category 7' and 'Other') were represented using stratified sampling.

The Random Forest was built using 200 decision trees (estimators) within the 80% training subset and was given a 'random_state' value of 42 to ensure reproducibility of the model. Reproducibility in this context is important due to – as previously mentioned – a vast number of features being present in the data.

3.3. Results and Performance Metrics

After training had been completed, the model's performance was evaluated on the remaining 20% of labelled data; the testing subset. This ended with the generation of the classification report, which provided us with the classification report, ultimately allowing us to gauge the success of this model on the testing subset.

In terms of accuracy, the report told us via a confusion matrix that there were very few misclassifications throughout the model implementation, meaning that the model correctly predicted a call as Category 6, Category 7, or 'Other' on the majority of occasions. Thus, we could deem this initial trial as accurate and precise.

In earlier sections of this report, we identified the priority metric to address as recall, as the heuristic-based approach had this at only 2%. From this initial testing, recall rate was as high as 0.985 (98.5%), showing a drastic improvement, and far surpassing our goal of 80%.

3.4. Predictions on Unlabelled Data

The final step of this supervised analysis saw the training model subsequently applied to all 2000 unlabelled call transcripts in the alternative dataset. After applying the same cleaning and transformation pipeline from the labelled data to the new dataset, the predictions made by the model were added into a new column, labelled 'Predicted_Category'.

Finally, these predictions were exported to a CSV file. This will allow whoever responsible for the business aspect of this task to refine their sales process and quality assurance measures accordingly. Ultimately the Random Forest method is a solid data-driven solution which can also be seen as a strong foundation for a further, semi-supervised machine-learning model.

4. Semi-Supervised Approach

4.1. Methodology

For the semi-structured model, the Random Forest Classifier was again demed the most appropriate approach. The aim was via a self-training model this time, to leverage the 2000 unlabelled examples in addition to the 400 labelled.

Once again, some slight pre-processing had to be completed prior to any analysis. To match with the labelled dataset, the column of 'Call_Type_Name' was dropped.

Following this, for each call in the unlabelled dataset, initial class probabilities were predicted using the existing supervised model. From here, pseudo-labels were assigned to class predictions with a probability equal to or over the confidence threshold, which was set at 95%. This step was taken to ensure that the model only learns from predictions that it is very confident about, therefore reducing noise and maintaining accuracy.

This pseudo-label data was then merged with the original training data to ultimately create an expanded training dataset. Subsequently, a new Random Forest Classifier was trained using this expanded dataset, with the same 80/20 training-test ratio as before. Predictions were made on this testing subset of the expanded dataset, similarly to the supervised model, before implementing the newly trained model on the full unlabelled dataset.

4.2. Evaluation and Results

From the unlabelled data, this model predicted 675 Category 7 calls and 57 Category 6 calls. While this may seem a little skewed, we must remember that in the labelled data there are significantly more Category 7 calls present.

In terms of model performance, as well as managing to maintain high accuracy, this model showed high Category 6 (92.5%) and Category 7 (100%) recall, potentially benefitting from the pseudo-labelling technique. However, the recall of the class 'Other' was significantly lower at 42%, showing that some misclassifications into Category 6 or 7 may have been present.

This was further investigated through a confusion matrix, which showed that 'Other' was in fact classified as 'Category 7' on 7 occasions. Nonetheless, this is not too alarming since Categories 6 and 7 recall was the priority of this task. Furthermore, the total average weighted recall falls just short of 80%, which is nearly achieving our pre-defined target, and is still a drastic improvement on the 2% recall of the heuristic-based approach.

4.3. Comparison with Supervised Model

Across all classes, the supervised model provided more stable and accurate predictions. Through identifying patterns from unlabelled data, the semi-supervised model did slightly improve Category 7 recall; however, this was at the expense of recall for the ‘Other’ category.

This may have been due to the model prioritising high recall, choosing to still include uncertain calls rather than missing them.

To refine this in future, perhaps we could increase the confidence threshold of the semi-supervised model to 98%, for example. We did not want to make this too stringent for this task; however, we are not ruling this out further down the line.

5. Conclusion and Future Implications

Throughout this project, both the supervised, and semi-supervised models managed to accurately recall Category 6 and Category 7 calls on the unlabelled call database, after being trained using labelled examples. Both models significantly outperformed the heuristic-based approach which the company had previously implemented.

One model however, outperformed the other. The supervised model managed to consistently recall calls belonging to categories out with Category 6 and Category 7, whereas the semi-supervised model tended to misclassify these as Category 7. Nonetheless, we believe that with more time and ability to fine-tune this, this model can be just as effective as the supervised model, which is – in our opinion – deployment-ready for business.

Appendix

Link to code - https://github.com/euansmith9/carfinance247/blob/755a655adb77f1091c5b970d2a6d7405ad4093d8/Project%20Notebook.ipynb