Methodology:

The training dataset consisted of a total of 14 different columns. Out of the 14 columns, three columns consisted of NULL values. The remaining columns either acted as an indexing column or an important feature column. Out of the 14 columns, "amt", "db\_cr\_cd", "payment\_category" , "payment\_category", "is\_international" columns were used for initial evaluation of traditional machine learning algorithms like Logistic Regression, Decision Trees, Naïve Bayes etc. The features were encoded using an Ordinal Encoder and null values for "db\_cr\_cd" was imputed by the “None” value. After processing the data and splitting it into a train and validation set, training was done on the traditional ML algorithms. All the traditional algorithms failed to produce impactful results. The highest multiclass accuracy achieved was 33 by Random Forest Classifier Algorithm. This analysis of these algorithms also concluded the feature importance for all the selected features was quite low.

The transaction description column on manual inspection hinted incorporation of useful information that can help categorize the transaction into one of the ten categories. The problem was not being treated as an NLP Text Classification Task. In the recent advancements in NLP research fields, transformers have proven to be state of the art for a variety of tasks, overcoming RNN and LSTM-based networks. Pre-trained transformers trained on Billions of documents have proven to provide better results when fine-tuned on a downstream NLP task. The text data in the transaction description column mainly consisted of financial terms. Therefore FinBERT, a variant of the BERT transformer model trained on a large financial corpus namely Financial PhraseBank. Using a variant of BERT trained on financial data brings the advantage of knowledge gathered from training on millions of financial documents.

After the choice of model and the architecture, the train data provided was processed to contain only “trans\_desc” and “Category”. The “Category” column was encoded with the help of a Label Encoder. Next, exploratory Data analysis was also performed on the processed data. Key findings from exploratory data analysis:

1. The categories are unequally distributed. Retain Trade and Entertainment Category consisted of more than 10k rows, whereas Education, Finance and Communication Services categories consisted of less than 500 rows. Figure 1 shows the distribution of Categories across the data.
2. The length of the transaction description follows a normal distribution. The majority of transaction descriptions consisted of 11-15 words with few outliers with 2 words minimum and 18 words maximum. Figure 2 shows the distribution of transaction description length across the data.

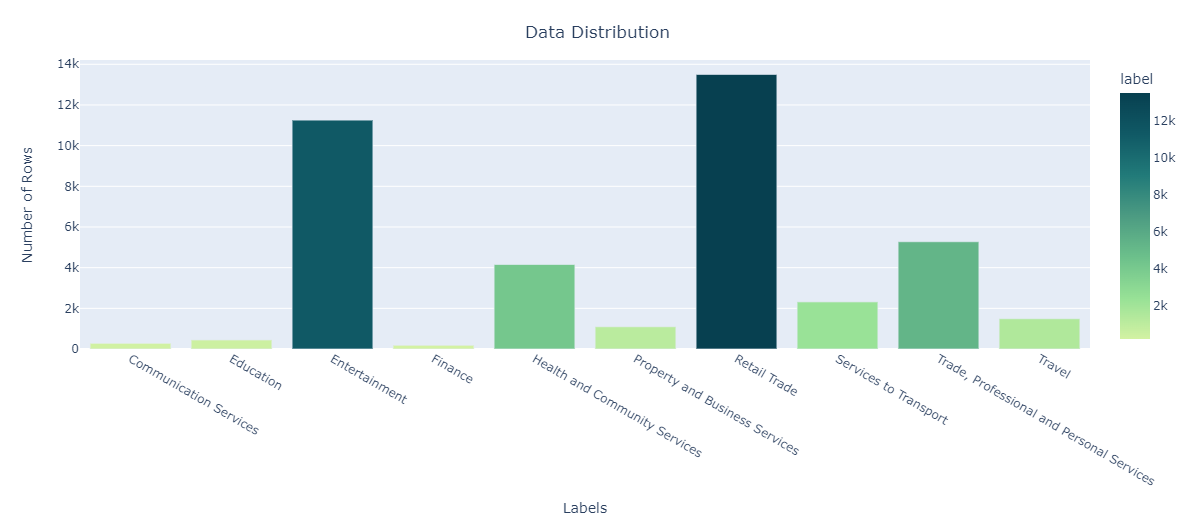


Fig1. Category Distribution

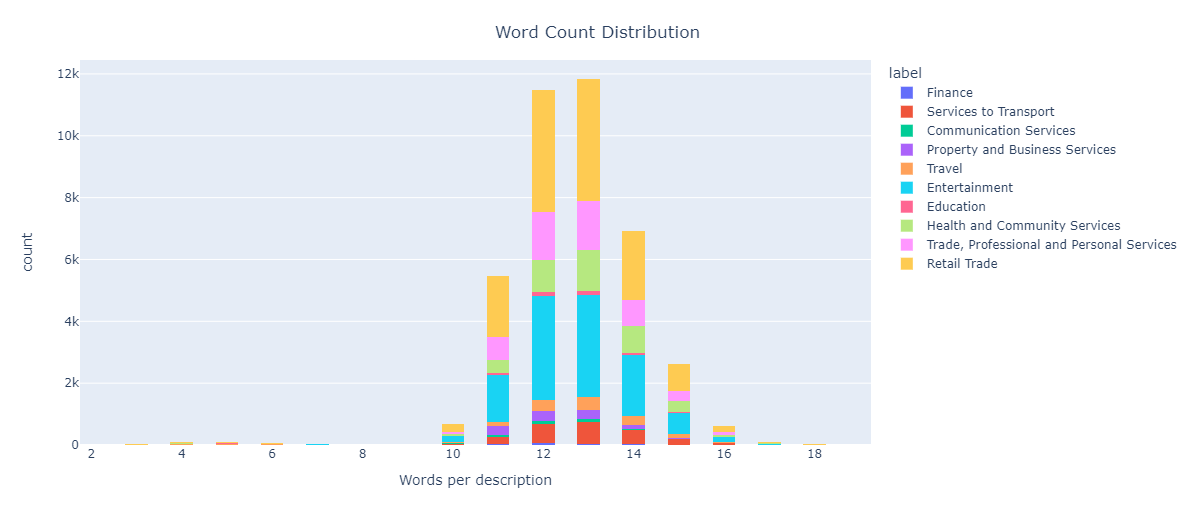


Fig2. Transaction Description Length (Word Count) Distribution

This data was then split into train and validation sets with a split ratio of 20%. The training and validation split consisted of 32000 and 8000 rows respectively. The transaction descriptions were tokenized with a max length of 32. The selection of the max length parameter was done on basis of length distribution obtained from the exploratory data analysis. All the text was processed to the lower casing as well. The tokenizer and model weights were obtained from the HuggingFace library. Huggingface library is an open-source library that provides weights for a variety of pretrained transformers.

After tokenising the data, FinBERT was trained on training data. AdamW optimizer with a learning rate of 0.00002 was used and the model was trained for 5 epochs. The model was evaluated on the validation set with a variety of classification metrics. The metrics included accuracy, precision, recall, f1 score and AUC roc score. After training for a total of 5 epochs on a batch size of 32, the model recorded a training loss of 0.7463. The evaluation result on the validation set are as follows:

* Loss: 0.955
* Accuracy: 0.6901
* F1 Score: 0.6855
* Precision: 0.6850
* Recall: 0.6901

To have a better look at the predictions made by the trained model, a confusion matrix and a multiclass roc AUC curve were also plotted. Fig 3 and Fig4 represent the confusion matrix and the Multiclass ROC AUC Curve.

.

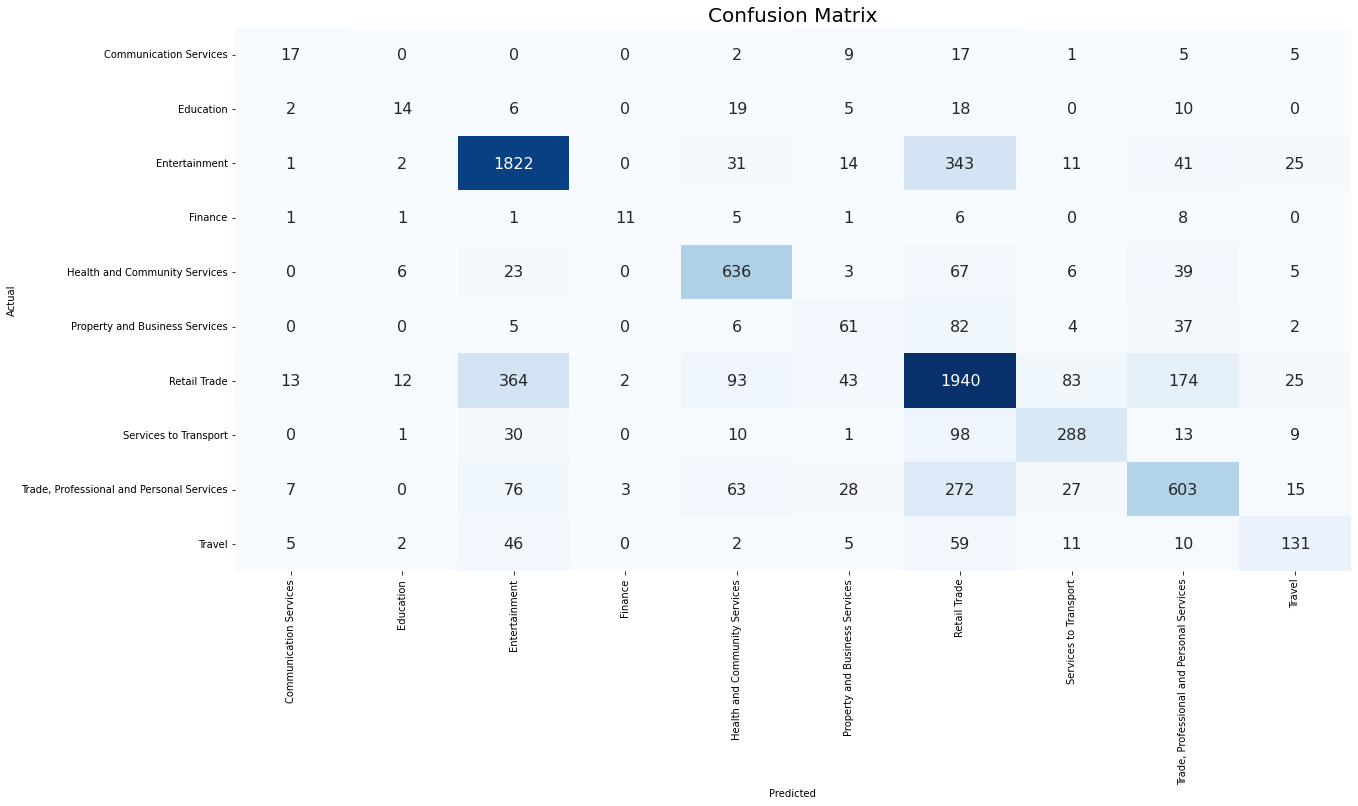


Fig3. Confusion Matrix

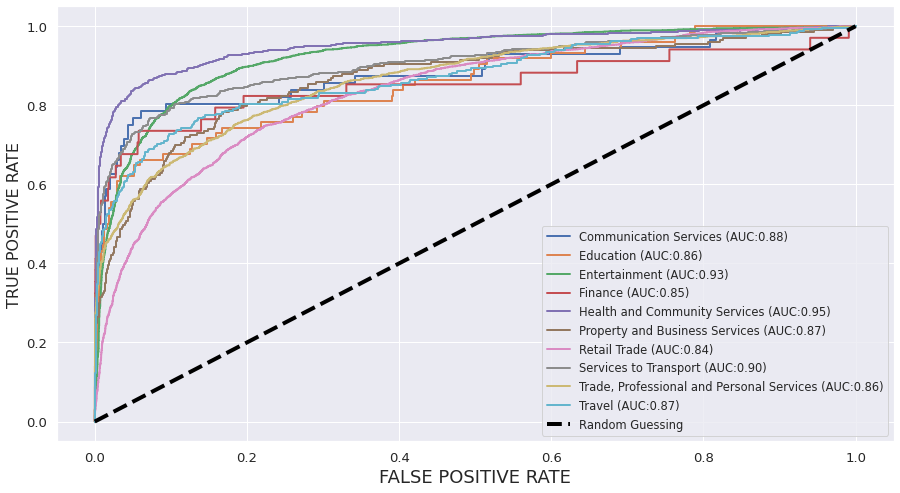


Fig4. Multiclass ROC-AUC Curve

At last, the inference was done on the test set provided. The test data was first loaded and processed in the same manner as the training data. Followed by processing, tokenization was done on the transaction description column. The tokenized data was passed on to the trained model for predictions. The predictions were then inverse transformed with the same encoder used for train data and replaced with the Category column of the test data. At last, the test data was saved as a Excel file.