Documentation

This work took me 3 days of coding and 2 days of documentation

Data analysis on the datasets and how it relates to the development of the model:

1. Understanding the Data:

Already given the explanations and data descriptions in two pdf files.

2. Data Cleaning:

Taking care of missing values ensures that the model is trained on accurate and comprehensive data. Consistency in data types is essential for both feature extraction and model input.

3. Exploratory Data Analysis (EDA):

Contributes to comprehending the distribution of transaction amounts and essential patterns. Detects anomalies and outliers in the data.

4. Feature Engineering:

Transaction trends can be captured by extracting components such as year, month, day, hour, and second.
Text descriptions can be preprocessed (lowercase, removing punctuation) and vectorisation using TF-IDF to turn text data into numerical features that can be used with machine learning models.
Numerical feature conversion from boolean user interests adds important information that can affect transaction categorization.

Reason why I choose this model:

A strong ensemble technique that performs well on classification tasks is Random Forest. It is perfect for this dataset with mixed features (transaction amounts, user interests, and text descriptions) since it can handle both numerical and categorical data.
Given the possibility for imbalance in transaction categories, Random Forest's ability to handle imbalanced datasets through methods like class weighting and sample algorithms is essential.

Reasoning and justifications:

1. Problem Understanding and Requirements:

The task is to predict the category of each transaction in the bank transactions dataset. The categories are predefined, making this a multi-class classification problem. The data includes numerical, categorical, and textual features, necessitating a model that can handle diverse types of input.

2. Selection of Techniques:

a. Data Preprocessing:
b. Model Selection:
c. Hyperparameter Tuning:

Predicting each transaction's category inside the bank transactions dataset is the task at hand. This categorization problem has multiple classes because the categories are predefined. The data includes textual, category, and numerical features, so a model that can handle a variety of input formats is required.

d. Model Evaluation:

Provides Accuracy, Precision, Recall, and F1-Score for each category to understand the model's performance on different classes.

- Accuracy: Measures overall correctness.
- Precision and Recall: Important for understanding the trade-offs between false positives and false negatives.
- F1-Score: Harmonic mean of precision and recall, providing a single metric that balances both concerns.

Steps to Develop the Model:

1. Load and Clean Data
2. Exploratory Data Analysis (EDA)
3. Feature Engineering
4. Model Training
5. Model Evaluation

Performance of the model and effectiveness:

The initial model's useful accuracy of 81% suggested that it could successfully classify transactions into the appropriate categories.
The marginal improvement in accuracy from 81% to 82% shows how well hyperparameter adjustment works to improve the model's predictive power and decision limitations. Even tiny increases in accuracy can have a big impact on the real world, especially in fields where precision is essential, like financial management, even though the change may only seem slight.
Users can feel more confident in the model's predictions with an accuracy of 82%. They can depend on the model to appropriately classify transactions, which will help them make smarter financial decisions. Fewer misclassifications translate into higher accuracy, which lowers the possibility of incorrect financial analysis or recommendations based on misclassified transactions.

	precision	recall	f1-score	support
ATM	0.97	1.00	0.99	109
Arts and Entertainment	0.86	0.86	0.86	7
Bank Fees	0.92	0.94	0.93	138
Check Deposit	1.00	1.00	1.00	5
Clothing and Accessories	0.67	0.36	0.47	61
Convenience Stores	0.71	0.85	0.77	373
Department Stores	0.50	0.28	0.36	43
Digital Entertainment	0.65	0.73	0.69	92
Food and Beverage Services	0.75	0.43	0.55	7
Gas Stations	0.65	0.68	0.67	254
Gyms and Fitness Centers	0.00	0.00	0.00	1
Healthcare	1.00	0.25	0.40	4
Insurance	0.74	0.64	0.69	36
Interest	0.75	0.50	0.60	6
Internal Account Transfer	0.98	0.98	0.98	242
Loans	0.88	0.84	0.86	410
Payment	1.00	1.00	1.00	2
Payroll	0.86	0.94	0.90	170
Restaurants	0.73	0.73	0.73	528
Service	0.55	0.32	0.40	19
Shops	0.69	0.59	0.64	148
Supermarkets and Groceries	0.74	0.73	0.73	324
Telecommunication Services	0.40	0.50	0.44	4
...				
accuracy			0.81	5171
macro avg	0.77	0.71	0.72	5171
weighted avg	0.81	0.81	0.81	5171

Figure1. Classification Report Before Hyperparameter

	precision	recall	f1-score	support
ATM	0.97	1.00	0.99	109
Arts and Entertainment	0.88	1.00	0.93	7
Bank Fees	0.93	0.93	0.93	138
Check Deposit	1.00	1.00	1.00	5
Clothing and Accessories	0.70	0.34	0.46	61
Convenience Stores	0.70	0.87	0.77	373
Department Stores	0.58	0.26	0.35	43
Digital Entertainment	0.68	0.71	0.69	92
Food and Beverage Services	0.50	0.14	0.22	7
Gas Stations	0.65	0.72	0.68	254
Gyms and Fitness Centers	0.00	0.00	0.00	1
Healthcare	1.00	0.25	0.40	4
Insurance	0.78	0.58	0.67	36
Interest	1.00	0.50	0.67	6
Internal Account Transfer	0.98	0.98	0.98	242
Loans	0.89	0.84	0.87	410
Payment	0.50	0.50	0.50	2
Payroll	0.85	0.93	0.89	170
Restaurants	0.74	0.76	0.75	528
Service	0.67	0.32	0.43	19
Shops	0.72	0.57	0.64	148
Supermarkets and Groceries	0.75	0.73	0.74	324
Telecommunication Services	0.67	0.50	0.57	4
...				
accuracy			0.82	5171
macro avg	0.78	0.68	0.71	5171
weighted avg	0.82	0.82	0.81	5171

Figure2. Classification Report Before Hyperparameter

Discuss potential ideas and future development plans:

When it comes to transaction category classification, the Random Forest model performed well. By exploring more complex models like Gradient Boosting or Neural Networks, as well as more complex feature engineering, the model's performance can be further improved.
We may collect and add more data over the following month to improve model generalisation. Improve text preprocessing and experiment with sophisticated text representation methods. Extensive hyperparameter adjustment and model selection are also possible.
Finally, we can implement the model for real-time prediction throughout the following 3 months. Create a feedback loop to enhance the model's performance over time and make it easier to handle more datasets.

References:

✔ "What Is Supervised Learning? | IBM," *Ibm.com*, Sep. 23, 2021. https://www.ibm.com/topics/supervised-learning.

✔ "What Is Random Forest? | IBM," *Ibm.com*, Oct. 20, 2021. https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

✔ Melanie, "The importance of Cross Validation," *Data Science Courses | DataScientest*, Sep. 29, 2023. https://datascientest.com/en/the-importance-of-cross-validation#:~:text=Cross%2DValidation%20is%20a%20method,to%20work%20on%20new%20data.

✔ "Hyperparameter tuning. Grid search and random search | Your Data Teacher," *Your Data Teacher*, May 19, 2021. https://www.yourdatateacher.com/2021/05/19/hyperparameter-tuning-grid-search-and-random-search/#:~:text=Grid%20search%20is%20the%20simplest,performance%20metrics%20using%20cross%2Dvalidation.