Deciphering Financial Documents

A Classification Journey

Giridharan M

09/06/2024

DECIPHERING FINANCIAL DOCUMENTS: A CLASSIFICATION JOURNEY

INTRODUCTION

This report outlines the approach, model selection, and results of a machine learning project aimed at classifying tables from financial statements into five categories: Income Statements, Balance Sheets, Cash Flows, Notes, and Others.

APPROACH

1. **Data Extraction and Preprocessing:**
    - HTML files containing tabular data were processed to extract relevant information.
    - Initial preprocessing steps included cleaning the extracted text to remove unnecessary characters and formatting.
2. **Feature Engineering:**
    - **Text Vectorization:** Text data underwent vectorization using TF-IDF (Term Frequency-Inverse Document Frequency) representation.
    - **Label Encoding:** Label encoding converted document labels into numerical format for model training.
    - **Word Count:** A feature representing the total number of words in the document was added.
    - **Word Cloud:** A visual representation of the most common words in the documents was created to understand the prevalent terms.
3. **Model Selection and Training:**
    - Various classification models were evaluated using the LazyClassifier library to identify the most suitable model.
    - The Support Vector Classifier (SVC) exhibited the best performance based on accuracy and generalization to new data.
4. **Model Evaluation:**
    - The trained SVC model was evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score.
    - Additionally, the ROC AUC score was calculated to assess the model's ability to distinguish between classes.

MODEL SELECTION

We opted for the **Support Vector Classifier (SVC**) due to its superior performance, particularly in accuracy and generalization. While models like LGBMClassifier and ExtraTreesClassifier showed higher training accuracies, their test accuracies were slightly lower. For instance, **LGBMClassifier** achieved **99.39%** training accuracy but only **92.05%** test accuracy, while **ExtraTreesClassifier** had **99.39%** training accuracy and **92.72%** test accuracy. Conversely, **SVC** had a slightly lower training accuracy at **96.85%** but maintained comparable test accuracy at **92.05%.**

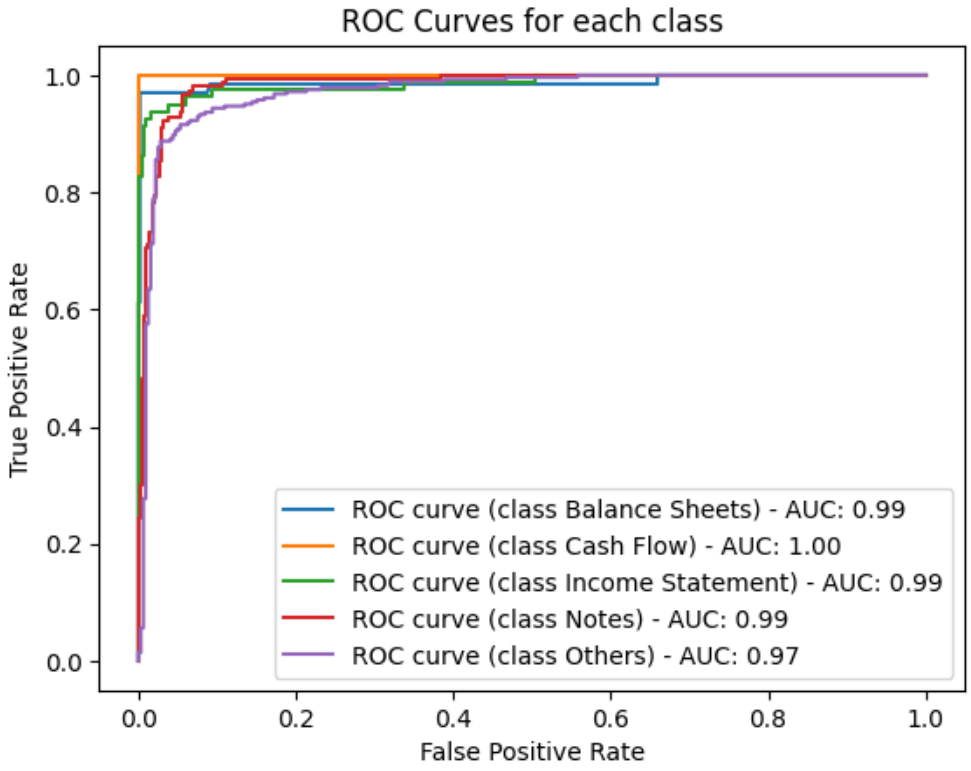
What distinguished SVC was its ability to generalize well to unseen data, evident in the narrow margin between its training and test accuracies. With a training accuracy of **96.85%** and test accuracy of **92.05%**, SVC showed **minimal risk of overfitting** and **superior generalization** compared to other models. Hence, owing to its balanced performance metrics and lower risk of overfitting, SVC emerged as the optimal choice for financial document classification.

RESULTS

1. **Accuracy:** The SVC model achieved an overall accuracy of **92.05%** on the test set, indicating precise classification of 92.05% of test samples.
2. **Precision, Recall, and F1-Score:** The average precision, recall, and F1-score were **93.41%, 93.78%,** and **93.56%,** respectively, demonstrating strong classification capabilities across all document types
3. **Classification Report (in %):**

Metric	Balance Sheets	Cash Flow	Income Statement	Notes	Others
Precision	98.48	93.75	93.67	90.07	91.10
Recall	97.01	100.00	91.36	87.18	93.33
F1-Score	97.74	96.77	92.50	88.60	92.20

4. **ROC AUC Score:** The ROC AUC score of **98.7%** indicated excellent ability to distinguish between different document types.

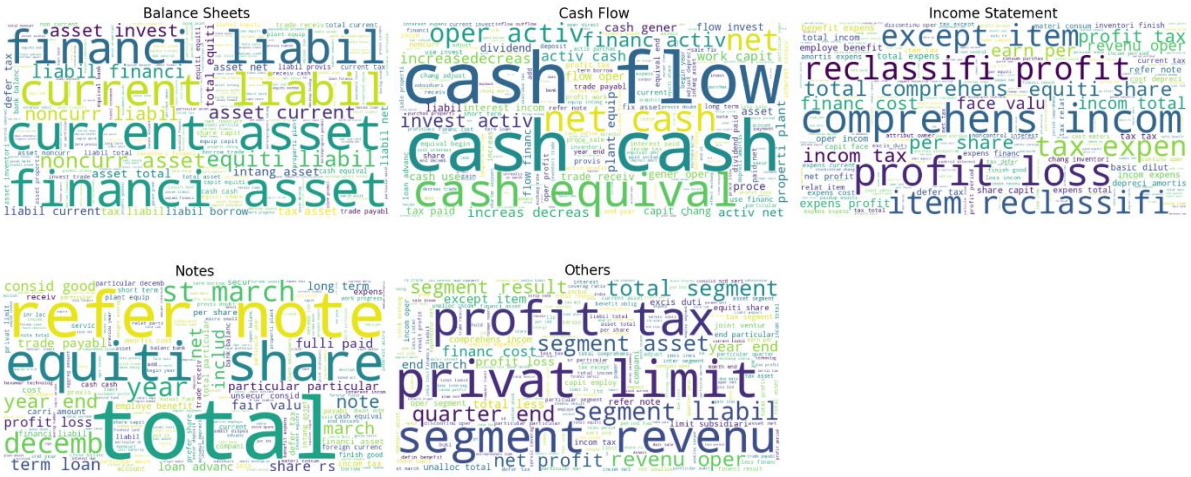


a. The model's exceptional ability to differentiate between different document types is evidenced by the ROC AUC scores for each class: Balance Sheets **(99%),** Cash Flow **(100%),** Income Statement **(99%),** Notes **(99%),** and Others **(97%).**
b. These scores highlight particularly high performance in classifying Cash Flow documents.
c. Additionally, the consistently high ROC AUC scores across all classes indicate robust discrimination capabilities, ensuring accurate classification across diverse financial document types.

CONCLUSION

The Support Vector Classifier (SVC) model achieved accurate classification of financial statement tables into their respective categories. With a balanced trade-off between accuracy, precision, recall, and generalization, the SVC model proves to be effective and reliable for this classification task. Future enhancements could involve fine-tuning the model parameters or exploring ensemble techniques to further improve performance.

WORD CLOUD:



In summary, This project demonstrates a robust approach to classifying financial statement documents, utilizing advanced feature engineering and machine learning techniques to achieve high accuracy and reliability in predictions.
