Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences

Code Blue: Can ML Save the Day?

C. Renner

Introduction

Project Overview and quick Look at the MIMIC- IV Dataset

Project Overview

- **Project Goal:**
  - Develop a predictive model to identify which factors influence patinet outcomes.

- **Models Used:**
  - Random Forest: Trained on 1100 + features extracted from dataset
  - Encoder Only LLM (BERT): Fine tuned on ICD titles for Classification

- **Dataset:** MIMIC- IV
- Focus of the Dataset:
  - ER and ICU patient data from 364k patients

Preprocessing and Feature Engineering

for Random Forest and BERT

Summary

- **BERT**
  - Aggregated ICD Titles, DRG Descriptions
  - Text Cleaning: Stop words removed, Lemmatization
    - Ineffective Approach
  - Tokenization: Wordpiece, Data Loader: Tensor input and target [MASK]
- **Random Forest**
  - **Feature Eng:**
    - One Hot Encoding
    - Binary Encoding
    - Ordinal Encoding
- Class imbalance removed

Random Forest

Results

Random Forest

- **Results:**
  - **Precision** 82-87% - significantly better than BERT
  - DRG severity and mortality parameters were the most important
  - ICD 9-10 Codes for DnR and Palliative care significantly increased mortality risk
  - Insurance Type was the 14th most important feature
- **Limitations:**
  - Bias in DnR and Palliative care
  - …

BERT

Feature Engineering and Model

Model dev with BERT

- **Model Definition:**
  - DistilBERT base uncased, 67M params
- **Training and Evaluation:**
  - Trained for 4 epochs ~2 Days duration
  - **Eval accuracy** 55%, **eval loss:** 69.2%

Tokenize

```
tokenizer = DistilBertTokenizer.from_pretrained('distilbert-base-uncased')
```

[6]

```
train_encodings = tokenizer(train_texts, truncation=True, padding=True, max_length=512)
val_encodings = tokenizer(val_texts, truncation=True, padding=True, max_length=512)
test_encodings = tokenizer(test_texts, truncation=True, padding=True, max_length=512)
```

[7]

```
balanced_df.shape
```

[4]

```
(10000, 3)
```

Data Loader/ Pipeline

```python
batch_size = 16

train_dataset = tf.data.Dataset.from_tensor_slices((
    {'input_ids': train_encodings['input_ids'], 'attention_mask': train_encodings['attention_mask']},
    train_labels
)).shuffle(len(train_texts)).batch(batch_size)
```

Compile and fit

```
# compile
optimizer = 'adam'
loss = SparseCategoricalCrossentropy(from_logits=True)
model.compile(optimizer=optimizer, loss=loss, metrics=['accuracy'])
```

```
#fit it
epochs = 7
history = model.fit(
    train_dataset,
    validation_data=val_dataset,
    epochs=epochs
)
```

Discussion & Conclusion

Discussion

- **Discussion:**
  - **Managing data size** was difficult -> multiple crashes and bugs
  - Feature Eng.: Selecting balanced and meaningful variables was difficult
  - BERT Text Processing: Lack of length normalization -> poor results for length normalization
  - **Selection of Features: BERT/RF**
    - RF: Numerical, Categorical Data --> high num of features handled well by RF
    - BERT: Handles Text Data better --> contextual encodings

Conclusion

- **Conclusion:**
  - RF: had **promising results** with 88% precision. Handled features effectively and results seemed very plausible.
  - BERT: 55% accuracy. Low Context length, text quality issues and imbalanced classes led to **poor result**.
  - Takeaway: Despite limitations and biases the project is a good proof of concept. Addressing the issues will lead to robust mortality prediction models.

Any Questions?

Sources:

- Kurpicz-Briki M. More than a Chatbot [Internet]. 2023. Available from: https://doi.org/10.1007/978-3-031-37690-0
- MIMIC-IV v3.1 [Internet]. 2024. Available from: https://physionet.org/content/mimiciv/3.1/
- MIMIC-IV documentation [Internet]. MIMIC. Available from: https://mimic.mit.edu/docs/iv/
- Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. Npj Digital Medicine [Internet]. 2018 May 4;1(1). Available from: https://www.nature.com/articles/s41746-018-0029-1
- distilbert/distilbert-base-uncased · Hugging Face [Internet]. 2024. Available from: https://huggingface.co/distilbert/distilbert-base-uncased
- RandomForestClassifier [Internet]. Scikit-learn. Available from: https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- StatQuest with Josh Starmer. Encoder-Only Transformers (like BERT) for RAG, Clearly Explained!!! [Internet]. YouTube. 2024. Available from: https://www.youtube.com/watch?v=GDN649X_acE
- GPT was used to streamline dev process.