SIOP Machine Learning Competition: team ___mifflin___

Ammar Ansari

Presentation Agenda

- Research process
- Initial problems + solutions
- What worked well
- What did not work
- Final submission overview

Research Process

- Consulted literature - previous SIOP Symposium talk referencing Hickman et. al. (2023): “Automatic scoring of speeded interpersonal assessment center exercises via machine learning: Initial psychometric evidence and practical guidelines”.
- Researched NLP techniques to gain familiarity (background statistical theory + Python libraries).

Initial Problems and Solutions: Missing Values

- For missing values in training set rating scores, imputed mean for each column.
- Missing values in text exercise questions were handled by concatenating all responses into a single column, then dividing all participants into groups based on which questions they received. These groups were then split into train and dev sets for NLP and ML analysis.

Initial Problems and Solutions: Cleaning Text

- Used stringr library in R to find and remove prompts left in participant responses.
- In order to fix misspelled words and correct phrases that were missing spaces between words, wrote a script that combined the pyenchant and wordninja Python libraries (see next slide for portion of script).

Example: The Power of Python NLP Libraries

```
for i in range(len(super_pub_df)):
    og_text_cell = super_pub_df.iloc[i,26]
    tokenized_og_text_cell = spacy_tokenizer(og_text_cell)

    for index, item in enumerate(tokenized_og_text_cell):
        if not dictionary_enchant.check(item):
            try:
                tokenized_og_text_cell[index] = dictionary_enchant.suggest(item)[0]
            except IndexError:
                wordninja_list = wordninja.split(item)
                joined_wordninja_list = " ".join(wordninja_list)
                tokenized_og_text_cell[index] = joined_wordninja_list

    super_pub_df.iloc[i,26] = " ".join(tokenized_og_text_cell)
```

What Worked Well

- Bag of n-grams (unigrams + bigrams)
- Term Frequency Inverse Document Frequency
- DistilBERT transformer model embeddings
- Cosine similarities between participant responses using each of the vectorization methods above
- Random Forests / Bagged Trees, depending on group
- Ridge Regression, depending on group

What Did Not Work Well

- SetFitTrainer
- LDA topic modeling
- TPOT automated ML pipelines
- LASSO regression
- K Nearest Neighbors

Final Submission Overview: Model 1

- Final submission used three models.
- For 5 out of the 7 ratings, used a model with the unigrams + bigrams cosine similarity scores and distilBERT embeddings cosine similarity scores as predictors.
- This model alone, using just n-grams and distilBERT scored a .491 mean R on the private leaderboard - combining it with the other 2 models increased it only .01, for the final score of .501.

Final Submission Overview: Model 2

- For rating_interprets_information predictions, used a model which additionally added the columns for cosine similarities of the TFIDF vectors for each group.

Final Submission Overview: Model 3

- For rating_involves_others predictions, used a different model that included the actual distilbert embeddings, as well as the TFIDF scores for each word instead of the cosine similarities of the TFIDF vectors.
- This model additionally included counts of the words "please" and "thank", as well as the overall and mean length of participant responses as predictors.

Final Submission Overview: Quantile Function

- Two sample Kolmogorov–Smirnov tests showed that while the vast majority of predictors between the training and test set came from the same distribution, the predictions (when rounding to the nearest value) did not.
- Using the quantile function in R to round predictions to have the same distribution as the outcomes on the train set improved model.

Takeaways

- Getting to know data is always important.
- Preparing and cleaning text greatly improved model.
- Simpler NLP techniques (bag of n-grams, TFIDF) can be effective.
- Transformer models are also very powerful.
- Learning / growth mindset!

Thank you!