Make Your Model Interactive

Alexander Baker

Finding Your Dataset

- First identify the question you would like to answer
- There are numerous websites that provide excellent datasets
- Some sites include:
  - Kaggle
  - Driven Data
  - Data.gov
  - UCI Data Repository

Understanding and Improving Your Dataset

- Understanding your data is a necessity for creating a good model
- Pandas is used to create and edit your dataframes
- Domain Understanding
  - Some preliminary research about the fields, data, and how it was collected may help you gain insight you would be unable to receive from only the data itself
- Exploratory Data Analysis
  - This practice is how insight is gained using the data
  - Matplotlib
  - Plotly
- Feature Engineering
  - This is the process of creating new fields that were not in your original dataset. These can be created with single field transformations, multiple field transformations, external datasets, etc.
  - These engineered features are driven by the insights found

Additional Data Improvements

- Handling your nulls
  - Incorporate sparse data with a different column
  - Delete the column
  - Delete the row
- Feature elimination
  - Principal Component Analysis
  - Recursive Feature Elimination
- Categorical Variables
  - One Hot Encoder
  - Label Encoder
- Standardization
  - Standard Scaler
  - Min Max Scaler

Model Creation

- Sklearn is an amazingly convenient and diverse library for ML
- ML Algorithm Suggestions:
  - Logistic Regression
  - Support Vector Machines
  - Decision Trees
  - Random Forest
  - XGBoost
- Separate into inputs and outputs
- Separate into training and testing sets
- Instantiate the model
- Train the model
- Predict values
- Evaluate outputs
- Further information can be found on how the algorithms are constructed and run

Validation

- K-fold Cross validation
  - Split into k number of sets
  - K-1 of these sets are used for training the remaining set is used for testing, then the output metrics are assessed
  - The set designated as the testing set is rotated until all sets were designated as a testing set once
  - This method we get multiple accuracies rather than just one
- Confusion Matrix
  - Breakdown of predicted, actual, type 1, and type 2 error
  - Allows detailed understanding of where the model needs assistance
- F1 Score
  - Precision - Ratio of correct predicted positive guesses over total positive guesses
  - Recall - Ratio of correct positive guess over the sum of true positives and false negatives
  - F1 score combines these metrics to assess a model
- ROC Curve
  - Graphs the true positive rate against the false positive rate (sensitivity and 1-specificity trade off)
  - This forms a curve and compares it to a random classifier
  - The closer the curve is to the top right corner the better the model performance

Tuning

- Each algorithm has its own set of parameters.
- Optimizing these parameters can decrease the chance of overfitting, and produce more accurate results
- The optimal parameters can be found using grid search
- Grid searching is the act of iterating through each set of proposed parameters, and assessing each one
- Highly inefficient, there are some optimized versions of grid searching algorithms you can implement to decrease the run time.

Dashboarding

- Saving the model
  - Models can be saved using pickling
  - Pickling sends your model to an external file that can be loaded into your dashboarding script
- The goal of this ML dashboard is to allow the users to create custom lines of inputs used to make predictions with
- Streamlit is an easy and intuitive library for python dashboarding
- Different dashboarding libraries include
  - Flask
  - Dash
  - Bokeh

Conclusion

- Each step of the basic ML pipelines was reviewed in this presentation
- Though this primer is simple, it can be used as a strong base for projects moving forward
- Try each step yourself
- There is still way more to learn

Moving Forward

- Using APIs, databases, and web scraping to collect your data
- Deep Learning
- Time series modeling
- Natural Language Processing
- Modeling using images
- Modeling using audio

Let’s Connect

- LinkedIn
- Illuminate AI
- Email: alex_b443@yahoo.com