Innovative. Technology. Partner.

Sharp Patient Risk Score POC – Technology Transfer Workshop

Analytics

Overview

- Housekeeping
- Review the project objectives
- Discuss the methodologies
- Discuss the result
- Demonstrate and discuss the project artifacts
  - Impala/Hive scripts for querying the Cerner tables in HDFS
  - Jupyter notebooks
    - Exploratory Data Analysis (EDA)
    - Data cleaning
    - Predictive model training and validation

Housekeeping

- You should have Anaconda (Python 2.7 version) already installed
- https://www.continuum.io/downloads
- The tech transfer directory and files are best viewed through a Jupyter notebook, which is easily obtained by installing Anaconda

- **Does anyone not have access to either the GitHub repository or the tech-transfer files?**

Project Review

- **Context** – Sharp HealthCare is a not-for-profit integrated regional health care delivery system in San Diego that consists of:
  - four acute-care hospitals
  - three specialty hospitals
  - two affiliated medical groups
  - a health plan

- **Goal** - to build a Proof of Concept implementation capable of identifying patients at risk for an adverse rapid response team (RRT) event

Methodologies

- Identify which patients had RRT events
- Identify patient features that are plausible causal factors for a sudden decline in health status
  - E.g. vital signs, medication usage, narcotics, movements between units, etc.
- Extract the subsets of data from which the desired patient features can be derived
- Process and analyze the features, while identifying which are sufficiently dense for model training
  - Select a cohort of counter-examples, i.e. Who did not have an RRT event?
- Model training and selection via cross validation

Results

- We were able to create probabilistic risk scores for patients
- Against validation dataset:
  - Accuracy = 80% (number correctly classified to their respective class)
  - Precision = 82% (positive predictive value)
- Performance improved by
  - Adding more training data, 0.5 years to 1.5 years of data
  - Adding more varied features
  - Further model optimization
- Performance impeded by
  - Lack of training instances
  - Lack of dense data features

Predictive signal definitely exists and our work suggests a clear path to improve performance

Project Artifacts

- The workflow
  - Use Impala to query Cerner data in Hive format
  - Conduct analysis and modeling using Anaconda in Jupyter notebooks
    - Mostly using Pandas and scikit-learn
- The structure of the project files
  - OVERVIEW.ipynb
  - etl-queries (directory)
  - notebooks (directory)
    - EDA (directory)
    - modeling (directory)

Demonstrate and Discuss

- Let’s see some code!


Questions