Pooja Rao

(She/her/hers)

- Postdoc, MSRI
- Previously: Stony Brook University
- Email: PRAO@msri.org

Research Field:

- Computational Fluid dynamic
- Quantum Computing

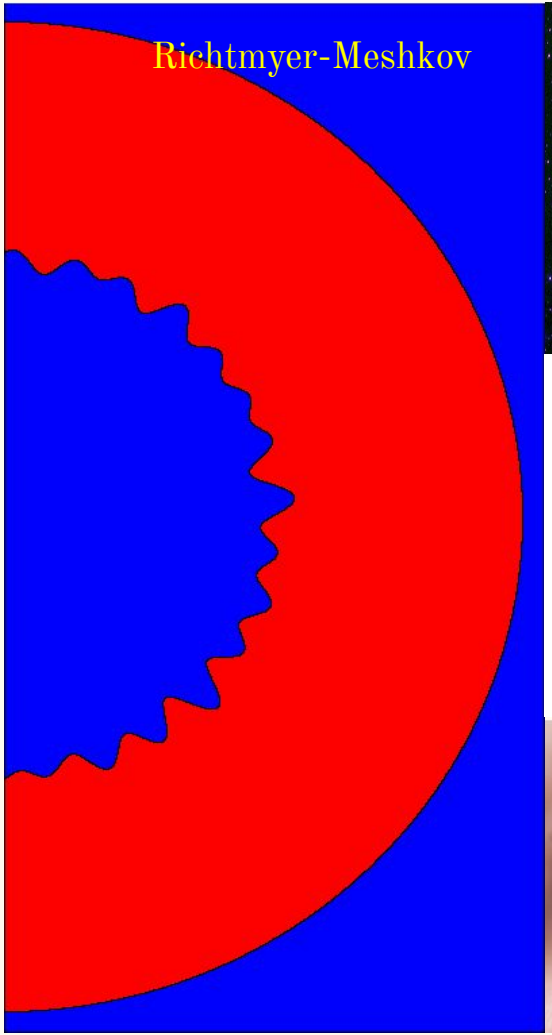
Quarantine Hobbies

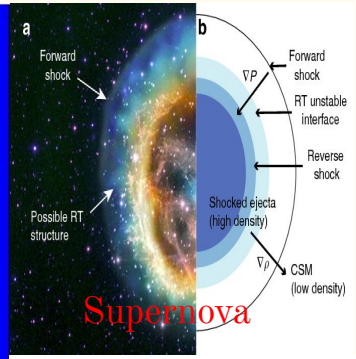
Fluid Dynamics

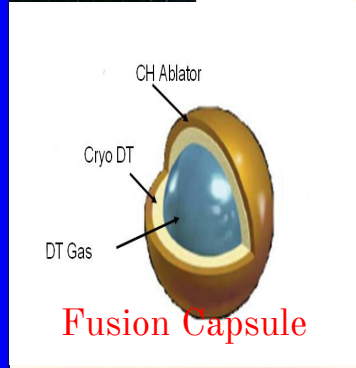
- Numerical modeling of fluids
- Turbulent mixing

Quantum Computing

- **Algorithms:** Search and solution counting algorithms
- **Applications:** Numerical integration, database search, linear equations.







Task

Dataset: Anonymized US Census Data for ~300,000 individuals.

Problem Description: Identify characteristics that are associated with person making more or less than $50,000.

Goal: Feature selection

- Pick out the “most important” features of the data.

Feature selection

Importance

- Selects the most relevant features
- Improves accuracy of the model
- More insight into the data
- Less training time

Three broad classes of methods

1. Filter methods (univariate) - fast, less accurate.
2. Wrapper methods - search through all the subsets, expensive.
3. Embedded methods - Random forest (consider smaller and smaller sets of features).

Approach

1. Explore the data
2. Pick a method
3. Pre-process the data
4. Implement the method
5. Analyze the results

Tools Used

- **Python and Jupyter notebook**

- **Pandas, Scikit-learn**

- **Code availability: https://github.com/poojarao8/uscensus_fs**

Data Exploration & Pre-processing in Pandas

- Missing headers for data columns
  - Hard to interpret
  - Parse headers from the metadata file
  - Type of data - both numerical and categorical

- Missing values marked by ‘?’ and ‘Not in Universe’

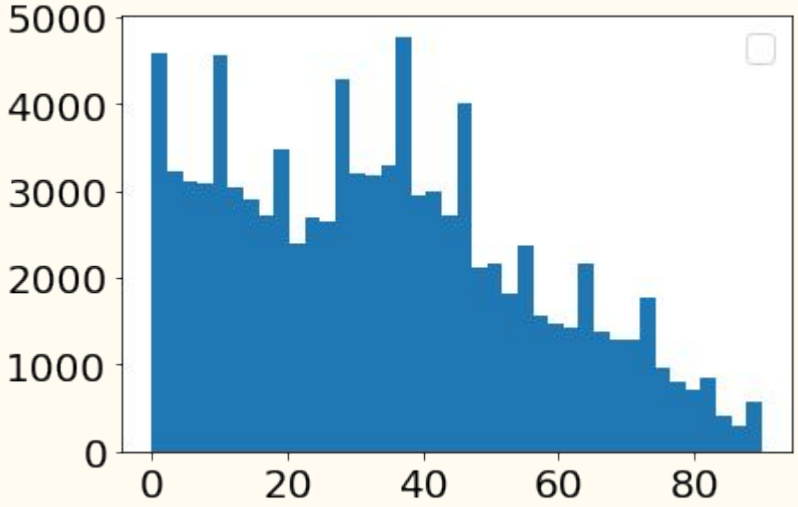
  - Deleting rows with ‘?’ gets rid of about 40% of the data

- Checked for outliers

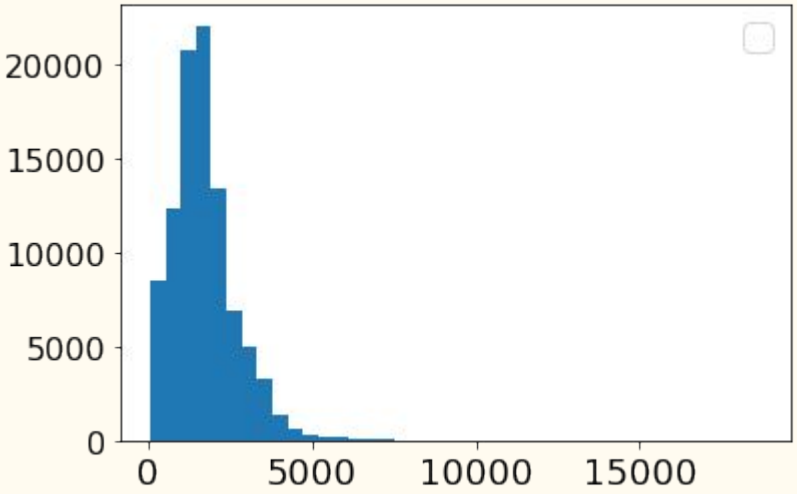
- Check against metadata

Preliminary Data Visualization

Age



Instance Weight



Picking an approach

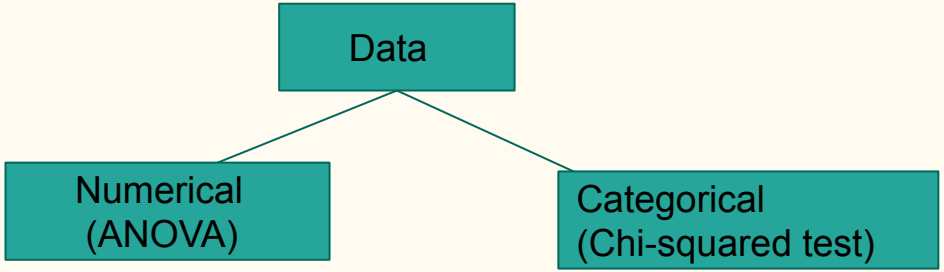
Goal: Feature selection

Numerical, categorical → categorical

- Pick different methods for each dataset.
- ANOVA for numerical and Chi squared for categorical

Pre-processing

- Separate the data (8 vs. 33).



- Encode the categorical data.
  - Change labels to numbers.
  - Sklearn's OriginalEncoder() and LabelEncoder() functions.

Basics and Terminology

Null-hypothesis: No relationship between instance and outcome.

Variance: How far is the data from the mean.

F-test: Ratio of variances

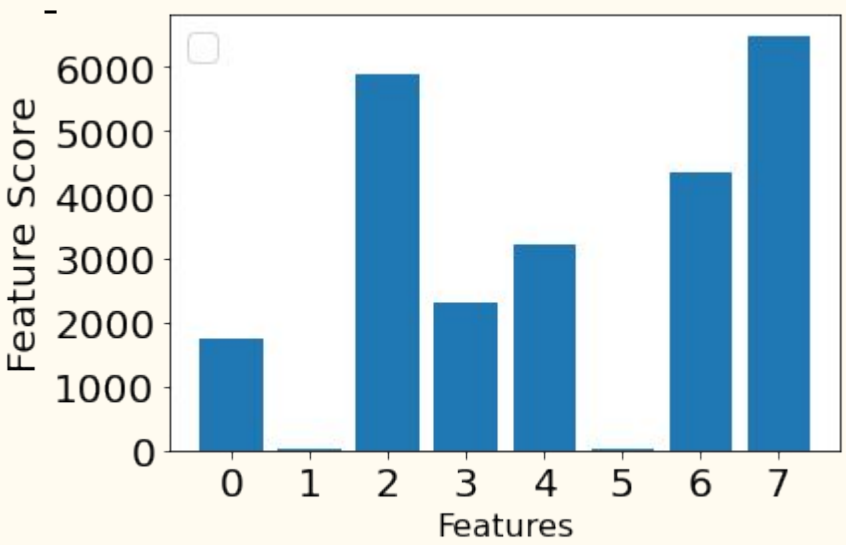
Contingency matrix: Cross-tabulated data.

P-values: Probability of getting the outcome you are getting given null hypothesis.

- Small p values implies relationship.

ANOVA for Numerical to Categorical

Analysis of variance (ANOVA)



Assumptions:

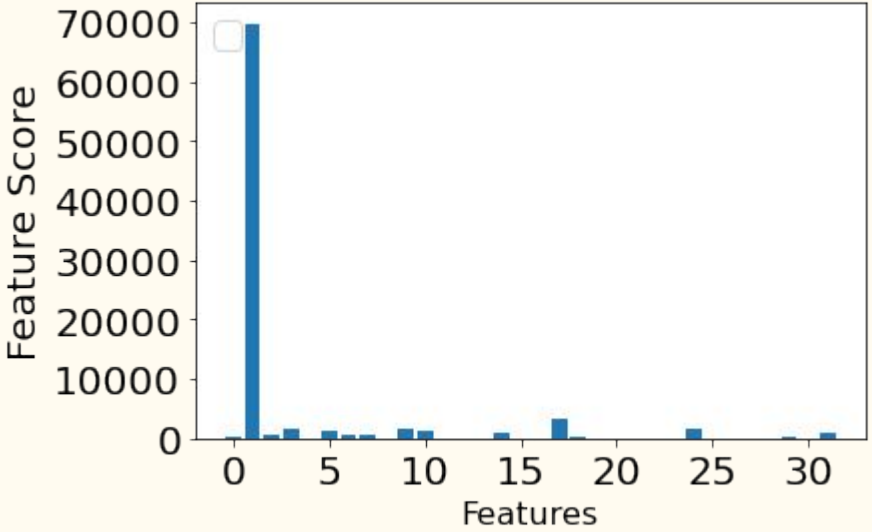
1. The data is normally distributed.
2. Observations are independent of each other.
3. Variance in each group is

Note:

1. If not normal, higher chance of false positives.

Chi-Squared for Numerical to Categorical

- Simple to implement
- Works for categorical data
- Uses contingency tables



Model Summary and Discussion

- Used filter based models for feature selection
  - Chi-squared and ANOVA

- Data cleaning and preparation
  - Missing values and headers
  - Data encoding

- Analysis
  - Industry code seems to be really important among the numerical data.
  - Chi-squared needs further looking into.

Further improvements

- Pre-processing: Normalizing, missing values, duplicates, outliers.
- Address the nan values in the data (due to deleting of rows that lead to very small values.)
- Mutiple features reduce overall accuracy
  - 0.95*0.95*0.95....
- Other approach - decision trees