Feature Selection Methods on Car Crash Data in 49 States

Zakk Loveall
Connor Kasarda

Presentation Organization

1. Introduction
2. Research
3. Methodology/Experiment
4. Results
5. Conclusion & Future Work

Introduction

Problem Statement

- ***Change in project direction…***
  - Original Approach
    - Feature selection and prediction of car crash severity between nations
  - Tried to reconcile/combine features from multiple datasets between countries
    - This proved to be infeasible give time allotted
    - Even if there were similar features, had different scalings
  - Found dataset that contained car accidents between states
    - Sobhan Moosavi. (2022). *US Accidents (2016 - 2021)* [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/3286750

- What's the best way to predict the severity of car crashes between states?
  - Factors to consider…
    - Feature Selection Techniques
    - Variance
    - Accuracy
    - Computation Time

Research

Current Research

- https://www.iihs.org/topics/fatality-statistics/detail/state-by-state
    - Fatal crash totals; deaths by road user; crash types; DUI; restraint use; rural versus urban
- R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh and A. A. Frefer, "Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity," *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*, Amman, Jordan, 2019, pp. 272-276, doi: 10.1109/JEEIT.2019.8717393.
    - Select factors and model for classifying severity (w/ respect to country, not state-wise)
- Zhang S, Khattak A, Matara CM, Hussain A, Farooq A (2022) Hybrid feature selection-based machine learning Classification system for the prediction of injury severity in single and multiple-vehicle accidents. PLoS ONE 17(2): e0262941. https://doi.org/10.1371/journal.pone.0262941
    - Use of feature selection, but no use of subsetting the dataset

Novel Approach

- Separate feature selection on subsets of dataset
  - Each dataset subset contains data from one state only
  - Predicated on the fact that there may be difference between states in the U.S.

- Try different strategies for said feature selection
  - Already existing feature selection methods
  - Variance Thresholding
  - Machine Learning Algorithm Components

Methodology/ Experiment

Sampling Method

- Stratified data based on states
- Randomly sampled 5000 points if larger than 5000
- Normalized the data
- Encoded it to account for strings
- Split into test and training data

Feature Selection Methods

- Selected top 10 features from each method...
  - Feature Selection Based:
    - *1.* Recursive Feature Elimination
    - *2.* Sequential Feature Selection
    - *3.* Variance Threshold
  - ML Algorithm Based:
    - *4.* Random Forest
    - *5.* Neural Network

- Cross validated on MLPClassifier
  - 3 Folds
  - Mean Scores Over Folds

What we Measured

- Ran each state through feature selection and calculated total mean
    - Accuracy score was utilized
- Calculated variance between selected features for each state
    - How different were the features between each state?
- Calculated time it took to run each selection method
    - Includes features selection, cross validations, variance calculations for each method
- In the python code, we tried to parallelize wherever possible
    - Typically involves setting ‘n_jobs’ to -1

Results

Sequential Selection

- Iteratively selects a feature subset and evaluates performance using a metric
- Adds or removes features from the subset until criteria is met
- Goal is to find the smallest (10) subset that has the best performance
  - Score: **80% Accurate**
  - Time Elapsed: **854 Seconds**
  - Variance: **0.75 Variation**

Recursive Selection

- Works similarly to sequential
- Works backwards when adding/removing features
- Recursively finds subsets
  - Score: **79% Accurate**
  - Time Elapsed: **325 Seconds**
  - Variance: **1.02 Variation**

Variance Threshold

- Used threshold of 0.01
  - If variance < 0.01, feature is not considered
  - Else, feature is kept
- Works by removing features with low variance
  - Score: **76% Accurate**
  - Time Elapsed: **147 Seconds**
  - Variance: **1.93 Variation**
    - **This value may need some review...**

Random Forest Selection

- Used random forest algorithm to measure importance of each feature
    - Gini Importance
        - Splitting based on features
- Importance scores are assigned to each feature based on performance
- Features with least importance are removed
    - Score: **76% Accurate**
    - Time Elapsed: **254 Seconds**
    - Variation: **0.93 Variation**

Neural Network Selection

- Uses neural network algorithm to select features
  - Use of permutation importance
    - Reshuffling values in each column
    - Re-evaluate model with shuffled feature
- Similar to random forest in the sense that it assigns importance
  - Score: **76% Accurate**
  - Time Elapsed: **226 Seconds**
  - Variation: **0.79**

Comparison

- Best time(s):
  - Variance Threshold (147 seconds)
  - Neural Network (226 seconds)
  - Random Forest (254 seconds)
- Most Variation:
  - Recursive Selection (1.02)
  - Random Forest (0.93)
  - Neural Network (0.79)
- Most Accurate:
  - Sequential Selection (80%)
  - Recursive Selection (79%)
  - Neural Network, Random Forest, Variance Threshold (76%)

Conclusion & Future Work

Conclusion

- Recommend Random Forest Selection for this specific problem
  - High Variability (0.93)
  - High Accuracy (76%)
  - Low Comp. Time (254 seconds)
  - Good tradeoff between computation time and accuracy, while not losing a lot of information
- Issues:
  - Intense dataset (a lot of samples)
  - Hard to find good predictors
  - Hard to find a direction

Future Work

- Neural Network and Random Forest faster, still lack a little in accuracy
    - Perhaps ML algorithms could be improved or simplified
        - Make it more accurate and keep speed of algorithm?
- Make predictions with varied ML algorithms
- Improve feature selection accuracy
    - Maybe investigate why certain features selected

References

- https://scikit-learn.org/stable/modules/classes.html
- https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents
- https://www.iihs.org/topics/fatality-statistics/detail/state-by-state
- https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0262941