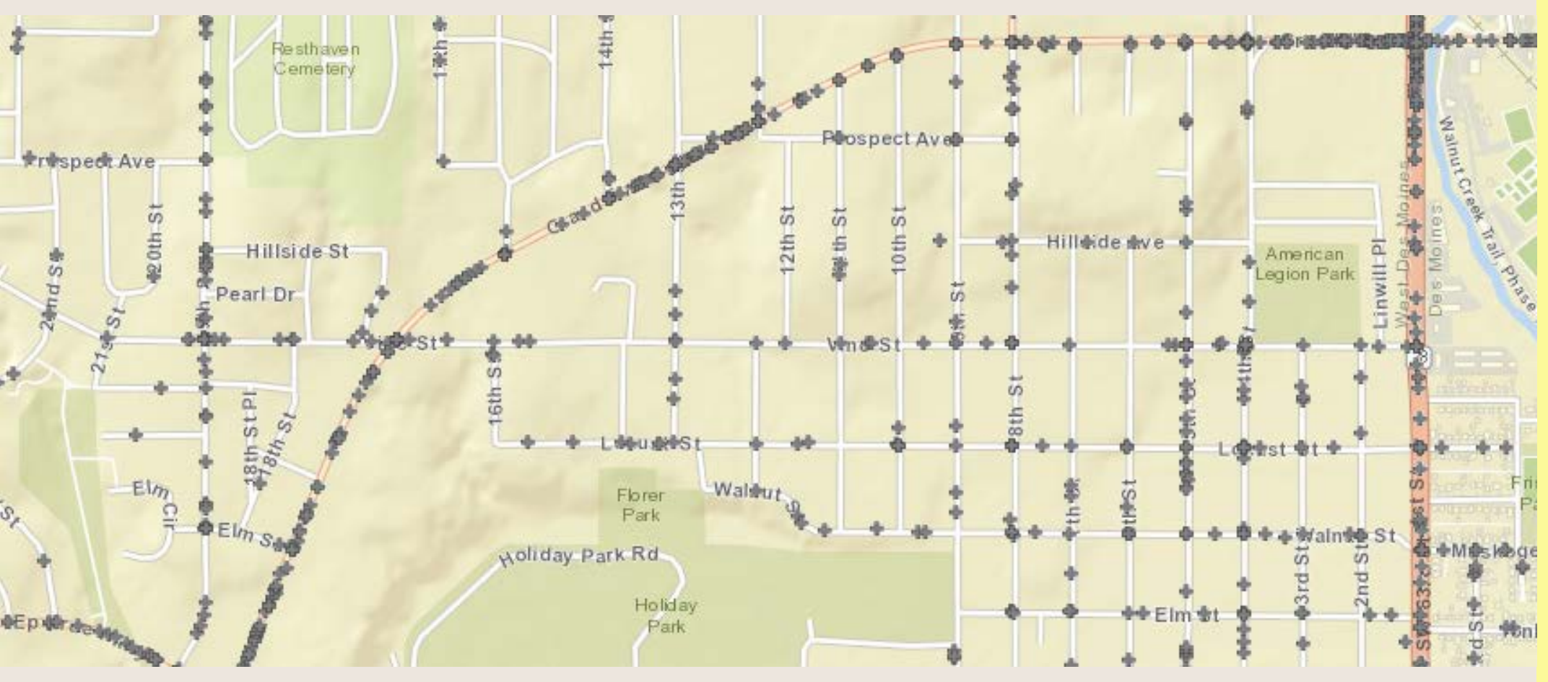
Alec Arshavsky

Reviving our Infrastructure to Save Lives

Introduction & Data

There is a lot of data on car accidents, but often local governments have limited resources to process it.

Most of the effort in improving infrastructure to reduce crashes is focused on intersections, but only 40% of accidents and 20% of fatal ones happen there.



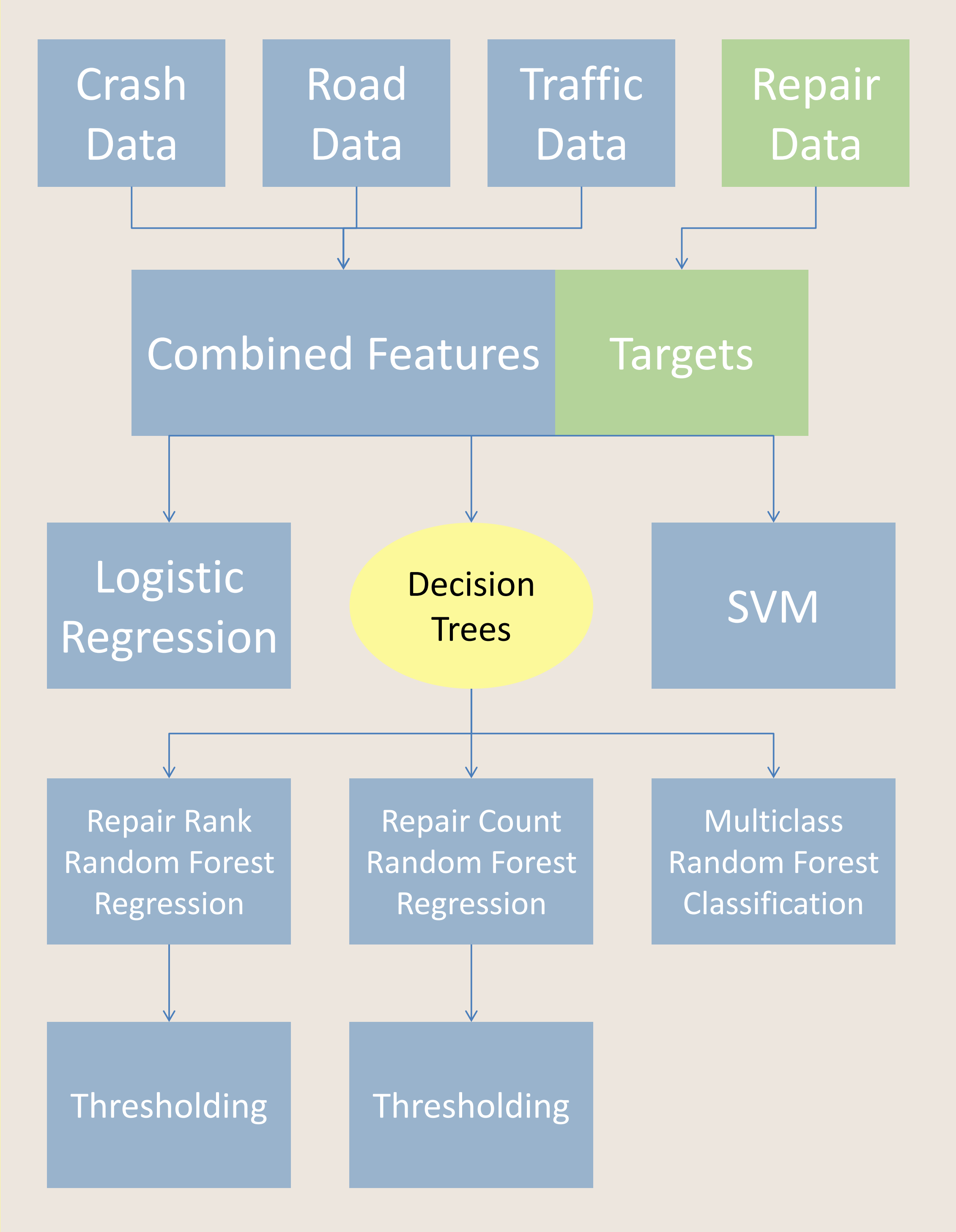
The goal of this project is to automate the suggestion process for road improvement candidates – a broader scope than efforts focused solely on intersections.

Selecting intersections to repair is a process that takes manual work as well as frequency analysis of crashes.

The dataset is a composite of four separate datasets from the Iowa DOT:

- Crash data
- Road features
- Traffic volume per road
- Iowa intersection improvement priorities

Features are road traffic volume, road length, number of lanes, width of dividing median, total fatalities, avg damage, avg alcohol level, avg severity, speed limit, and the number of crashes on the road.



Methods

Logistic Regression

- Good for large datasets, robust, efficient

SVM

- Good for large datasets, can handle noisy datasets, can handle non-linear feature division

Random Forest Decision Trees

- Can handle non-linear features, can perform multiclass classification as well as regression, resilient in noisy datasets
- Used three kinds of tree:
  - Regression based on intersection priority ranking
  - Multiclass classification of discretized classes based on priority ranking
  - Regression based on number of intersection candidates on a given road

Thresholding

- Convert regression data into binary classification by dividing regression estimates across threshold

Results & Conclusions

Logistic Regression

Accuracy:	92.47%
False Positives:	68.09%
False Negatives:	91.77%

SVM

Accuracy:	86.46%
False Positives:	89.92%
False Negatives:	87.79%

Repair Rank Forest

Accuracy:	89.48%
False Positives:	67.36%
False Negatives:	50.34%

Repair Count Forest

Accuracy:	93.57%
False Positives:	38.12%
False Negatives:	82.85%

Multiclass Forest

Accuracy:	93.54%
False Positives:	17.20%
False Negatives:	89.44%

Estimated

Actual

9844	8	5	1	1	1
117	23	2	1	1	0
118	1	21	0	0	0
151	2	2	4	0	0
119	0	1	0	11	1
147	3	2	0	0	2

Confusion Matrix for the Multiclass Random Forest

This Matrix shows the results for the six evaluated classes of intersection improvement candidates. The first category represents the roads that were not on the list, and the next five categories represent classes based on the improvement priority ranking

False Positives

- A potential sign of further improvement candidate roads
- Too many is indicative of error in the algorithm rather than further candidates

False Negatives

- Highly prominent in all of the algorithms applied, Most road classification is biased towards not needing improvement.
- Likely due to the incongruence between comparing roads and specific intersection-based improvement recommendations

Discussion

- The decision tree algorithms outperformed logistic regression and SVM in terms of false positives and negatives.
- Applying a metric of intersection candidates inherently creates a noisy dataset, making it harder to evaluate error without further examining erroneous results
- There are many ways to balance false positives and negatives when absolute accuracy is of lesser importance

Acknowledgements

I would like to thank the CS 229 teaching staff, Professors Ng and Duchi, and, in particular, my project TA Nihit for the help that they provided in advancing this project. I would also like to thank the Department of Transportation of Iowa for compiling extensive data which proved extremely helpful to the success of this project.

Future Work

- Unsupervised Algorithms
  - May detect patterns not present within intersection ranking markers, creating a more robust algorithm for suggesting problematic zones
- Further investigation of “false positive” results, which may be indicators of road improvement necessity, should be conducted to determine costs and subjective danger
- Further work with decision trees to fine-tune algorithm based on these findings can be done, both to decrease false negatives and keep some “false positive” potential candidates