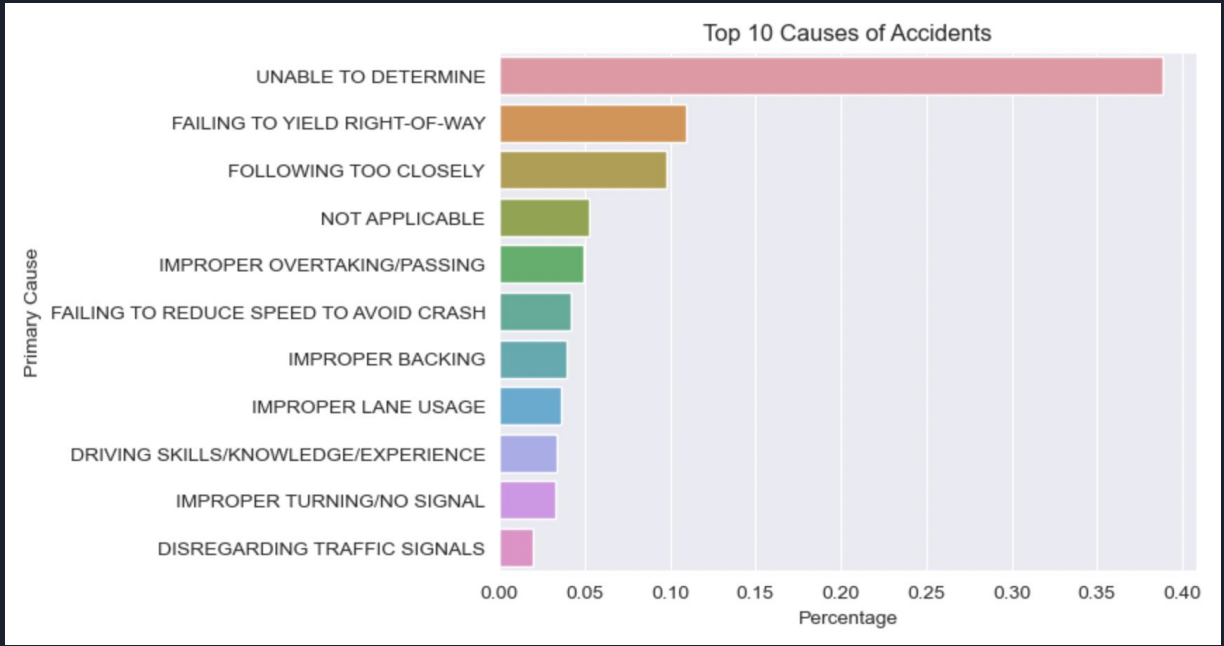
Predictive Insights into Car Collision Causes

By Nechama, Juan, and Colby

Business Problem

- Determine the primary cause of car accidents in the city of Chicago
- Build a classification model that will aid us in this task



Data Driven Solution

Data Source

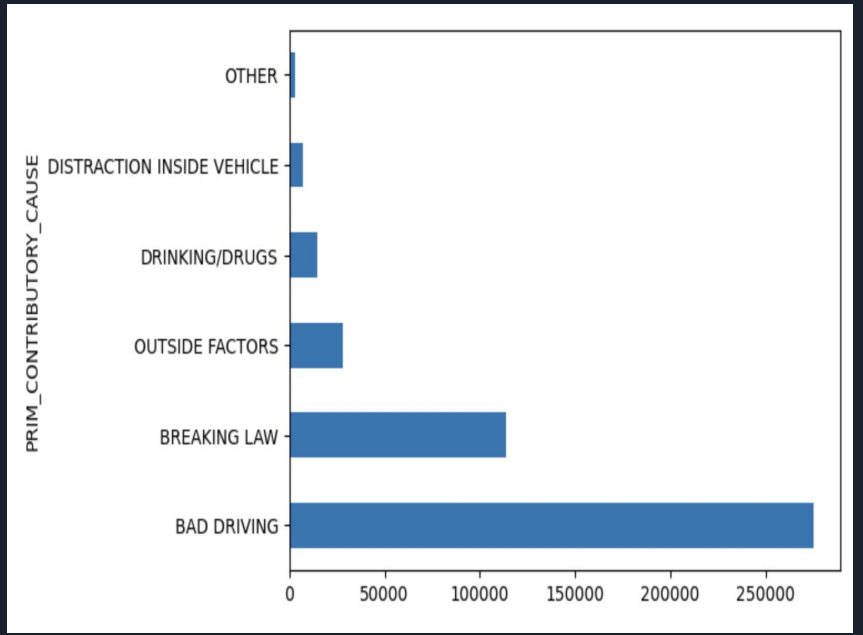
- Sourced data from Chicago Government on car crashes
- Dropped crashes with unknown ‘Primary Cause of Crash’
- Data from 2015 - present

Data Analysis

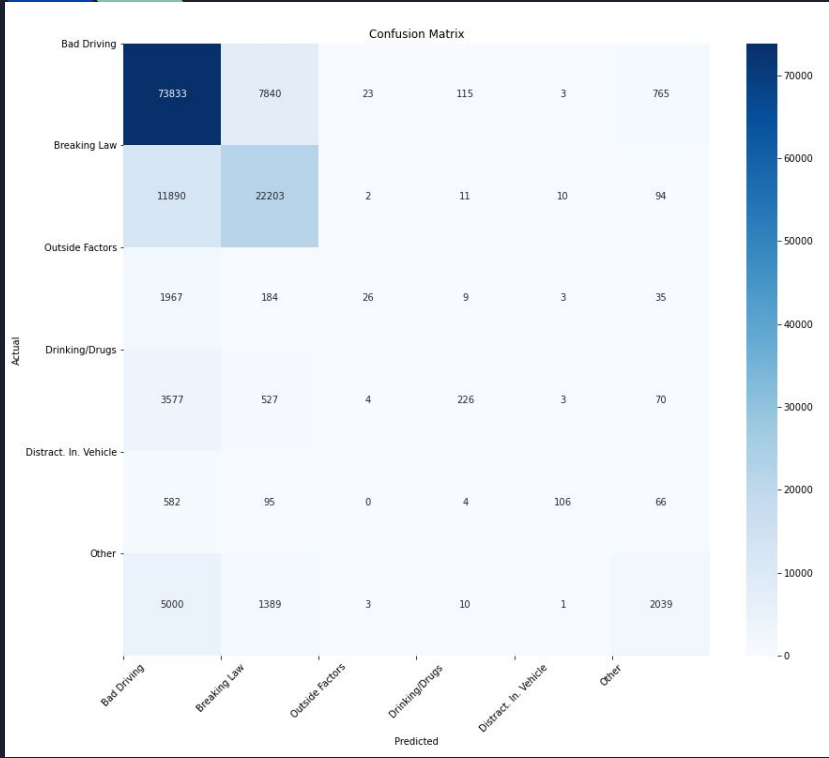
- Created a three baseline machine learning multi class classifiers
  - KNN model
  - **Random Forest** (Winner!)
  - Logistic Regression
- Tuned and adjusted models in order to boost performance

Simplifying the Target

- Started with 40 primary causes of accidents
- Binned into 6 relevant categories
  - **Bad Driving**
    - Ex: Following to closely, improper lane usage
  - **Breaking Law**
    - Ex: disregarding the stop signs, failing to yield right of way
  - **Outside Factors**
    - Ex: Construction, Visibility, Weather
  - **Drinking/Drugs**
  - **Distraction Inside Vehicle**
    - Ex: Texting, Distraction Inside Vehicle
  - **Other**



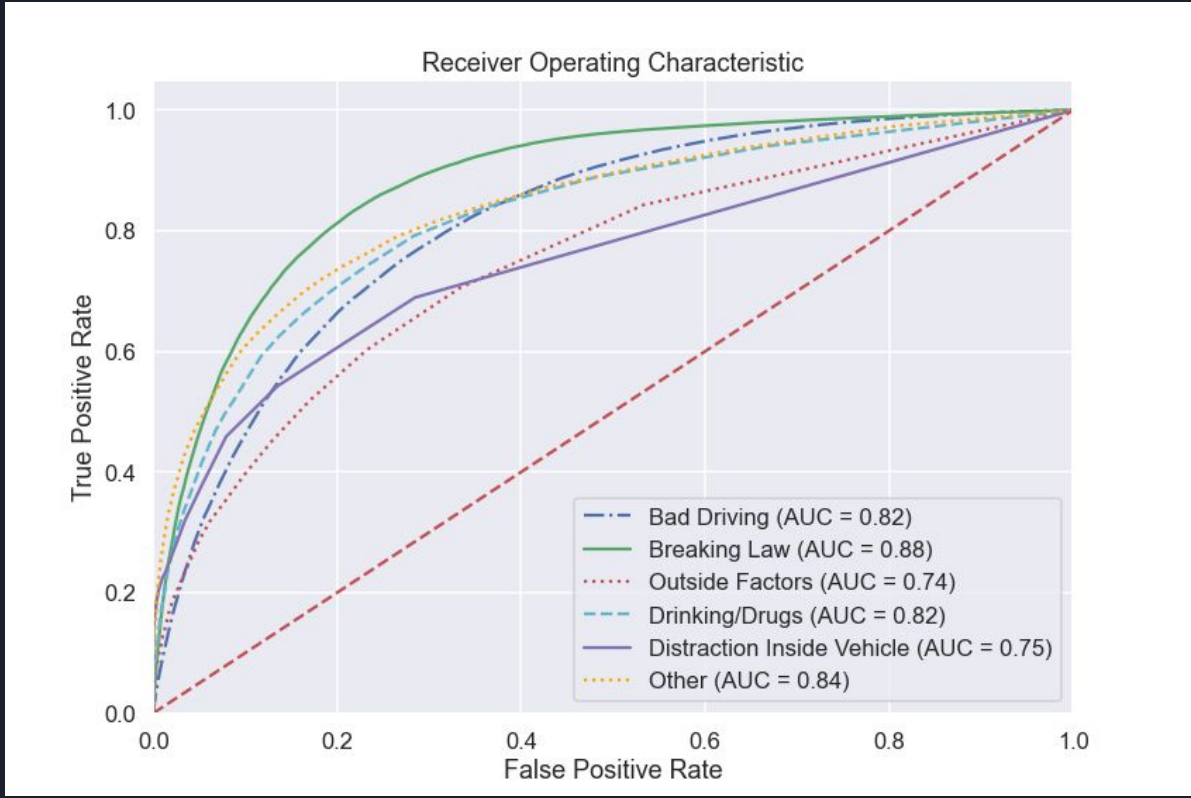
Random Forests Classifier



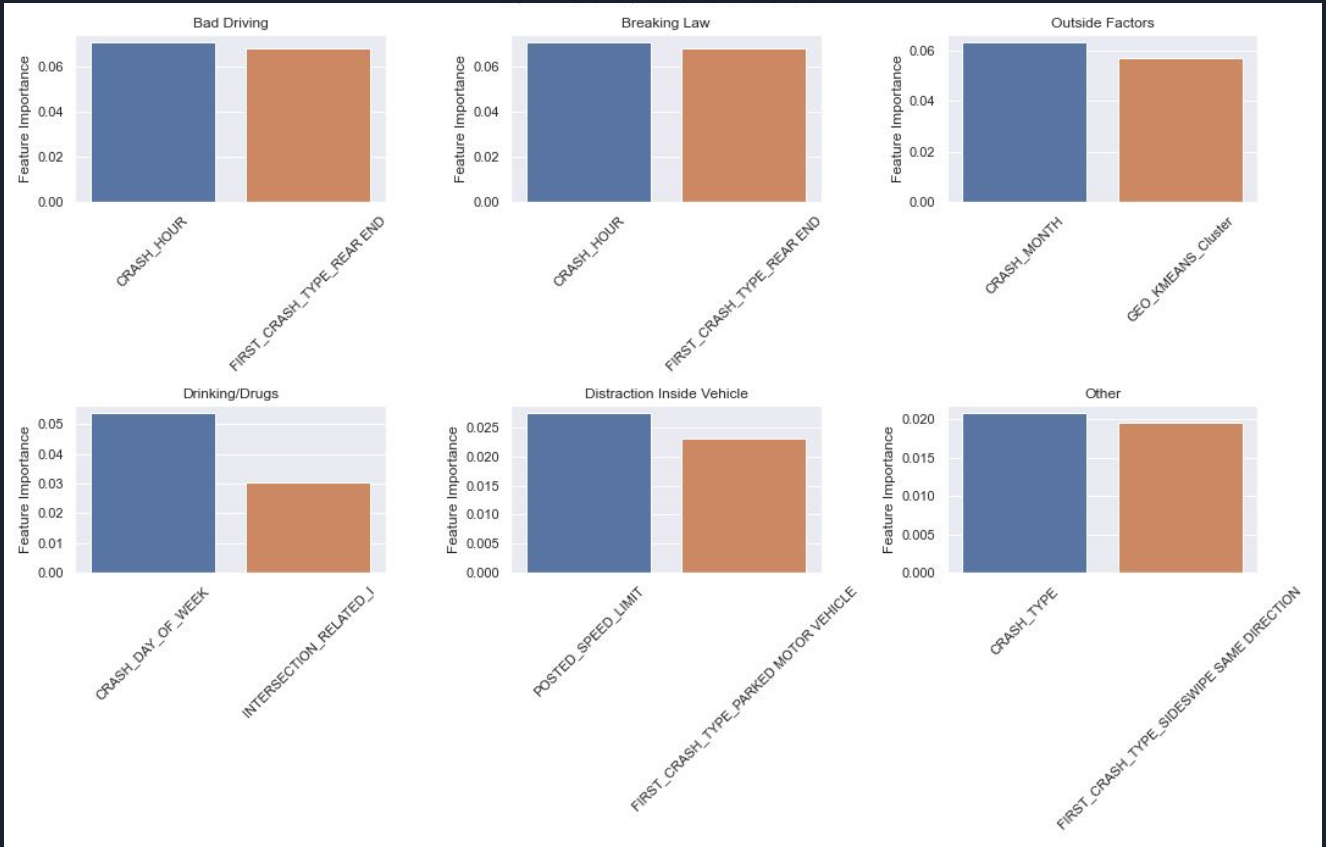
- Accuracy: ~74%

	Bad Driving	Breaking Law	Outside Factors	Drinking /drugs	Distraction	Other
Prec	76%	69%	48%	61%	85%	68%
Recall	90%	64%	1%	6%	11%	25%
F1 Score	82%	67%	3%	10%	20%	36%

ROC-AUC Curves



Top 2 Important Features for Each Class



Recommendations

- Our model provides a helpful tool for any investigation into a car crash where the primary cause of crash is unknown
  - Police
  - Insurance

- Prioritize Important Features in Investigation

Next Step

- Is our model generalizable to other cities?
- Test model on the most current data from the City of Chicago?
- Find other important features while trimming the features we are already using
  - Age and Sex of driver

Any Questions?