Colorado Oil & Gas Spill Analysis

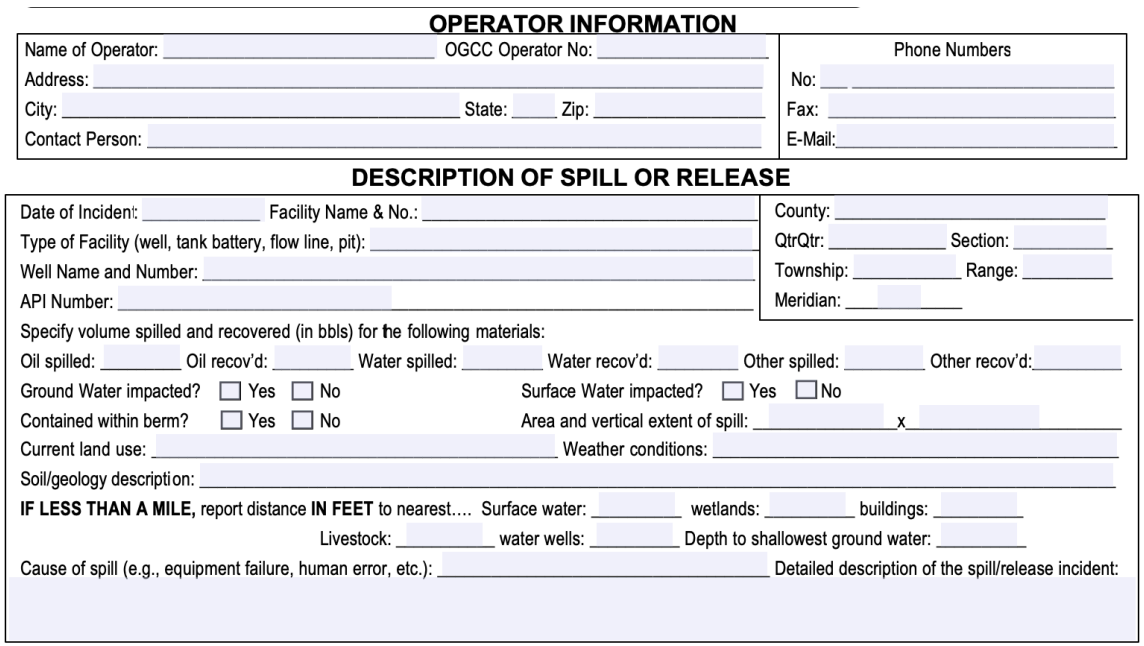
Alexander Parker
August 7th, 2019

Question & Dataset

Key Points

- ~~Question: Can we predict when and which wells will have a spill?~~
- Question 2.0: What factors contribute to large vs. small spills?
- Pros:
  - Spill-specific features
  - Class balance
  - Non-well locations
- Cons:
  - No predictive utility
  - Spill severity vs spill occurrence

COGCC Form 19 Snippet



OPERATOR INFORMATION

Name of Operator: _____ OGCC Operator No: _____
Address: _____
City: _____ State: ____ Zip: _____
Contact Person: _____

Phone Numbers
No: ___ _____
Fax: _____
E-Mail: _____

DESCRIPTION OF SPILL OR RELEASE

Date of Incident: _____ Facility Name & No.: _____
Type of Facility (well, tank battery, flow line, pit): _____
Well Name and Number: _____
API Number: _____

County: _____
QtrQtr: _____ Section: _____
Township: _____ Range: _____
Meridian: ____ ____

Specify volume spilled and recovered (in bbls) for the following materials:
Oil spilled: _____ Oil recov'd: _____ Water spilled: _____ Water recov'd: _____ Other spilled: _____ Other recov'd: _____
Ground Water impacted? ☐ Yes ☐ No Surface Water impacted? ☐ Yes ☐ No
Contained within berm? ☐ Yes ☐ No Area and vertical extent of spill: _____ x _____
Current land use: _____ Weather conditions: _____
Soil/geology description: _____
IF LESS THAN A MILE, report distance **IN FEET** to nearest.... Surface water: _____ wetlands: _____ buildings: _____
Livestock: _____ water wells: _____ Depth to shallowest ground water: _____
Cause of spill (e.g., equipment failure, human error, etc.): _____ Detailed description of the spill/release incident:

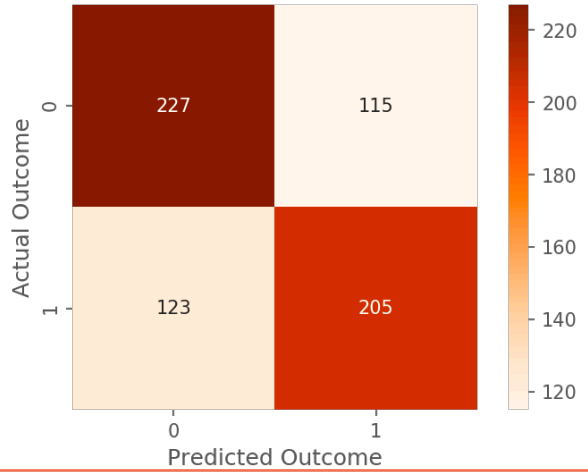
Model Discussion

Logistic Regression Model

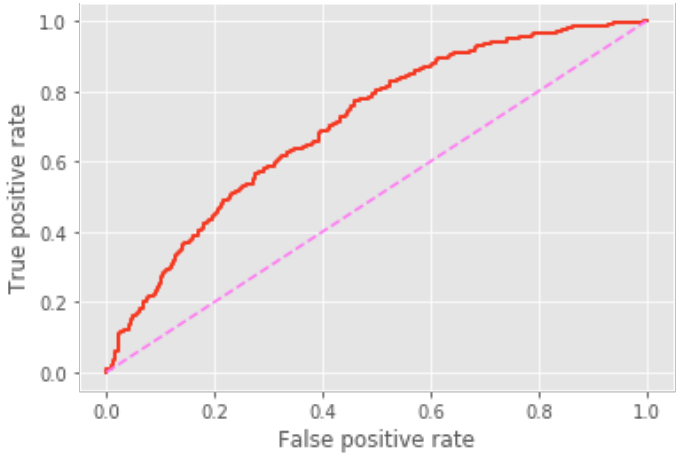
Key Points

- Logistic model selected based on ROC AUC score
  - 5-Fold CV, 5 times
- Random forest model performed on par or better to logistic model
  - Significant overfitting problem
  - Ensembling would reduce interpretability
- Test set scoring:
  - ROC AUC: 0.71
  - Accuracy: 0.64
  - Precision: 0.64
  - Recall: 0.63
  - F1 Score: 0.63

Confusion Matrix



ROC Curve



Feature Analysis

Feature Subcategories

Feature	Probability
historical_unknown	0.137
facility_type_GAS GATHERING PIPELINE SYSTEM	0.235
waste_spilled	0.242
public_byway	0.270
condensate_spilled	0.273
waters_of_the_state	0.303
facility_type_PARTIALLY-BURIED VESSEL	0.319
surface_owner_FEE	0.359
facility_type_PIPELINE	0.363
oil_spilled	0.376
equipment_failure	0.384
current_land_use_CROP LAND	0.387
facility_type_WELL	0.400
facility_type_FLOWLINE	0.405
season_spring	0.425
surface_owner_STATE	0.439
facility_type_GAS COMPRESSOR STATION	0.440
season_summer	0.443
facility_type_PIT	0.455
other	0.458
facility_type_GAS PROCESSING PLANT	0.462
days_from_first	0.467
current_land_use_OTHER	0.477
surface_owner_OTHER	0.481
facility_type_CENTRALIZED EP WASTE MGMT FAC	0.482
facility_type_OTHER	0.494
season_winter	0.500
surface_owner_TRIBAL	0.502
current_land_use_NON-CROP LAND	0.505
facility_type_GAS GATHERING SYSTEM	0.512
water_spilled	0.512
residence_/_occupied_structure	0.517
livestock	0.520
flowback_spilled	0.521
surface_owner_FEDERAL	0.524
facility_type_OIL AND GAS LOCATION	0.539
facility_type_PRODUCED WATER TRANSFER SYSTEM	0.556
facility_type_OFF-LOCATION FLOWLINE	0.562
facility_type_TANK BATTERY	0.570
facility_type_WELL PAD	0.570
facility_type_WATER GATHERING SYSTEM/LINE	0.583
surface_water_supply_area	0.589
drill_fluid_spilled	0.595
other_spilled	0.615
human_error	0.739

0.474 Train Dataset Y=1 Frequency

Facility Type Features

Feature	Probability
Gas Gathering Pipeline System	0.235
Partially-Buried Vessel	0.319
Pipeline	0.363
Well	0.400
Flowline	0.405
Gas Compressor Station	0.440
Gas Processing Plant	0.462
Centralized Ep Waste Mgmt Fac	0.482
Other	0.494
Gas Gathering System	0.512
Oil And Gas Location	0.539
Produced Water Transfer System	0.556
Off-Location Flowline	0.562
Tank Battery	0.570
Well Pad	0.570
Water Gathering System/Line	0.583

Spill Cause Features

Feature	Probability
historical_unknown	0.137
equipment_failure	0.384
other	0.458
human_error	0.739

Feature Analysis (Cont'd)

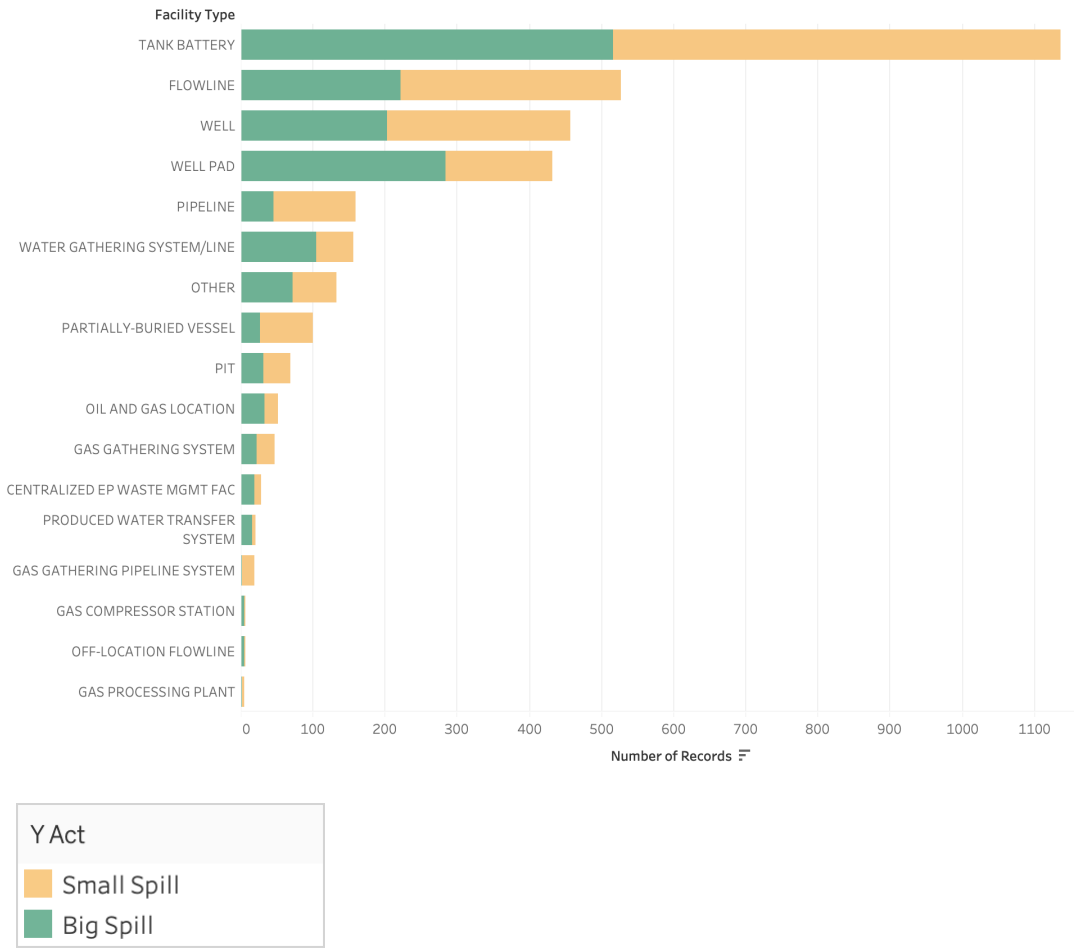
Facility Type Features

Key Points

- Boolean feature for each facility type
- Gas-related infrastructure is less likely to create large spills vs. water systems

Facility Type Features	
Feature	**Probability**
Gas Gathering Pipeline System	0.235
Partially-Buried Vessel	0.319
Pipeline	0.363
Well	0.400
Flowline	0.405
Gas Compressor Station	0.440
Gas Processing Plant	0.462
Centralized Ep Waste Mgmt Fac	0.482
Other	0.494
Gas Gathering System	0.512
Oil And Gas Location	0.539
Produced Water Transfer System	0.556
Off-Location Flowline	0.562
Tank Battery	0.570
Well Pad	0.570
Water Gathering System/Line	0.583

Spill Classification by Facility Type



Feature Analysis (Cont'd)

Spill Cause Features

Key Points

- Human error is one of the best predictors for if a spill will be large
- Limited number of observations (3k unknown) makes this feature less useful

Spill Cause Features

Feature	Probability
historical_unknown	0.137
equipment_failure	0.384
other	0.458
human_error	0.739

Spill Descriptions

"**Truck driver drove off with the hose still attached to the tank.** This ripped the entire pump and system out of the ground and the truck dragged it with it for roughly 15-20 yards before stopping. Water began back-flowing out of the off-load bay transfer system"

"…when the driver experienced a bloody nose. While the **driver was reaching for a roll of paper towels in the cab…both the truck and pup trailer [left] the road**, tip over and come to rest on the passenger side against a power pole"

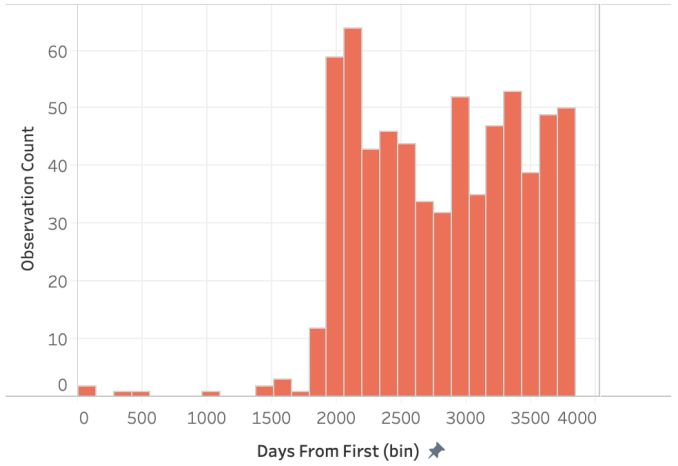
Logistic Regression Prediction Breakdown

Tableau Dashboard of Test Set Predictions & Respective Features

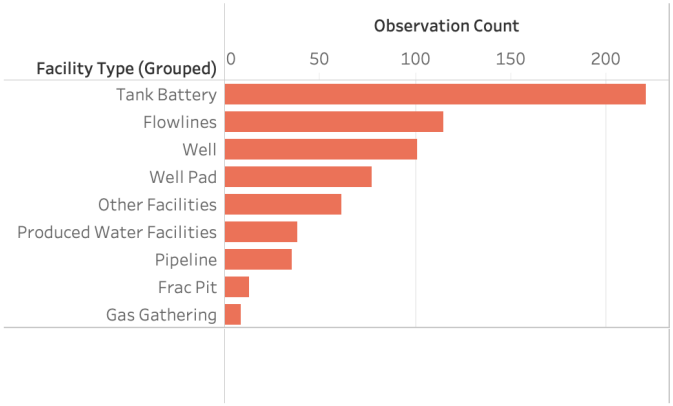
Confusion Matrix

	Y Act	
Y Pred	Small Spill	Big Spill
Small Spill	227	123
Big Spill	115	205

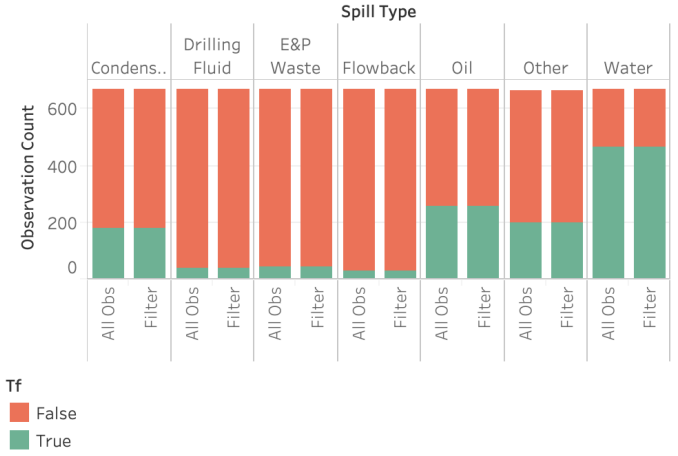
Observations by Days Elapsed



Observations by Facility Type



Spilt Chemicals Breakdown



Conclusion & Future Work

Conclusion

- Producing wells have less spill risk than related oil and gas activities
- Human error is a significant indicator for larger spills, but a more complete dataset would be useful for confirmation
- Produced water transport is another cause of large spills
  - Confirmed by by facility type and spilt chemical type

Future Work

- Time series data
  - Spills occur around bouts of activity rather than specific locations
  - Predict no spills versus spills
- Spill cause NLP
  - Working theory: big spills mostly caused by transport trucks
- Do certain companies have more human error incidents

Questions