Crash Data Prediction

of Delaware Counties

Group 3: Soha Ahmed, Shantanu Deshpande,
Kylie Dickinson, Steven Ha, Thinh Nguyen

Data & Target Variable

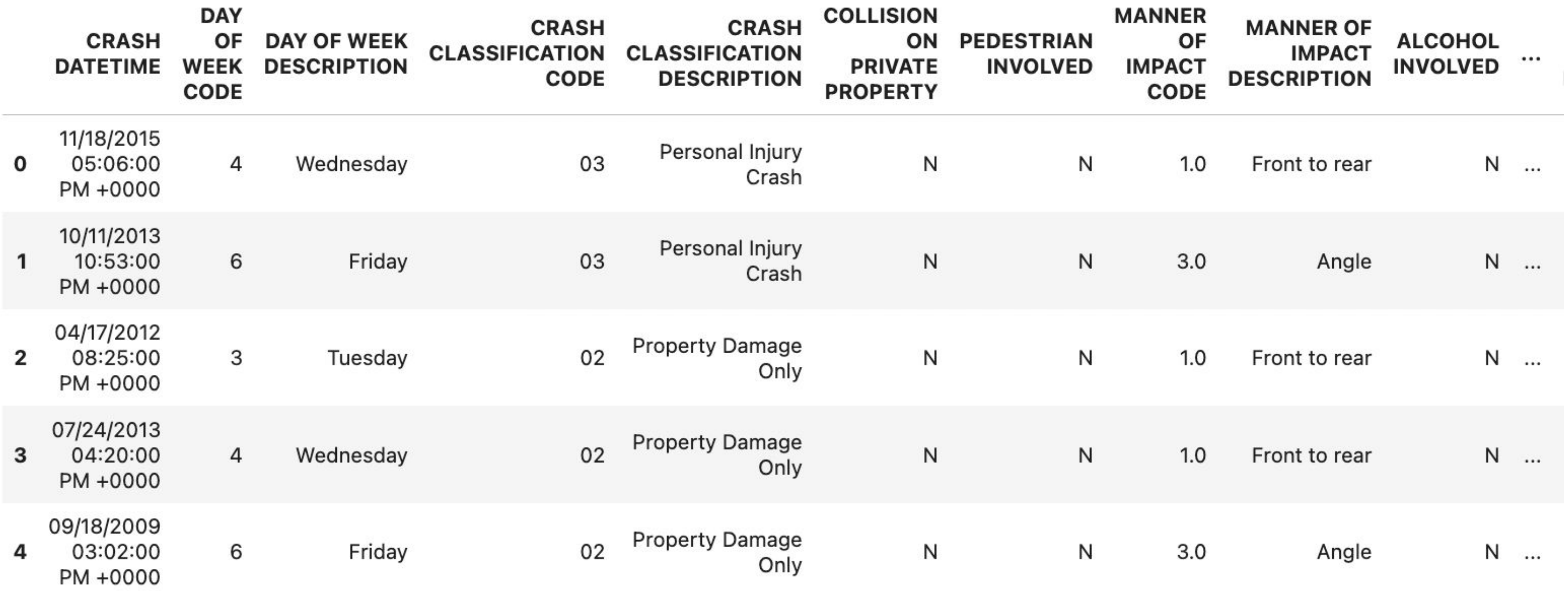
- Data contains 39 columns; 531, 115 entries of an accident
- We want to predict the Crash Classification Code/Description, categorical variables, in the data:
  - 01 – Non-Reportable
  - 02 – Property Damage Only
  - 03 – Personal Injury
  - 04 – Fatality Crash

GOAL:

Predict crash severity using known conditions at the time of the accident to support public safety, policy, infrastructure, education and autonomous systems

Data Source: Delaware Department of Safety and Homeland Security (DSHS)

Dataset Description



	CRASH DATETIME	DAY OF WEEK CODE	DAY OF WEEK DESCRIPTION	CRASH CLASSIFICATION CODE	CRASH CLASSIFICATION DESCRIPTION	COLLISION ON PRIVATE PROPERTY	PEDESTRIAN INVOLVED	MANNER OF IMPACT CODE	MANNER OF IMPACT DESCRIPTION	ALCOHOL INVOLVED	...
0	11/18/2015 05:06:00 PM +0000	4	Wednesday	03	Personal Injury Crash	N	N	1.0	Front to rear	N	...
1	10/11/2013 10:53:00 PM +0000	6	Friday	03	Personal Injury Crash	N	N	3.0	Angle	N	...
2	04/17/2012 08:25:00 PM +0000	3	Tuesday	02	Property Damage Only	N	N	1.0	Front to rear	N	...
3	07/24/2013 04:20:00 PM +0000	4	Wednesday	02	Property Damage Only	N	N	1.0	Front to rear	N	...
4	09/18/2009 03:02:00 PM +0000	6	Friday	02	Property Damage Only	N	N	3.0	Angle	N	...

Data Cleaning Methodology

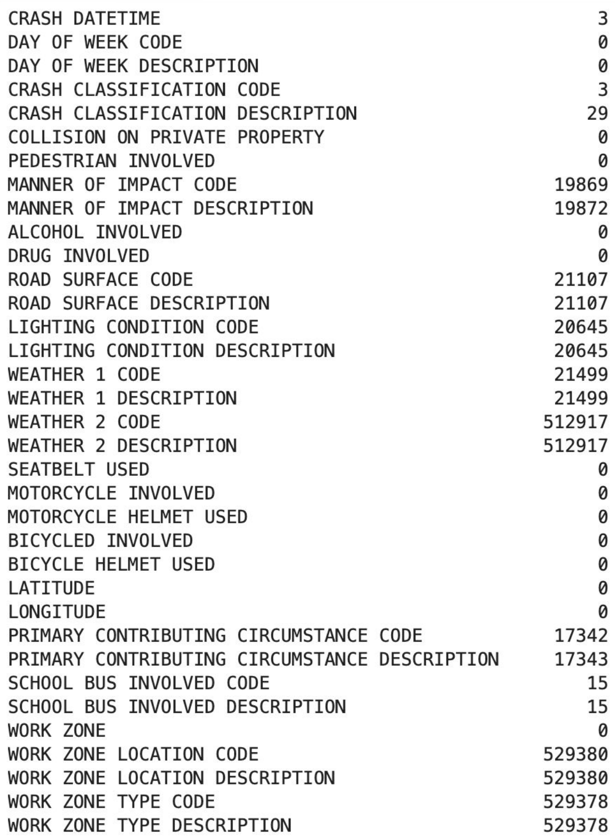
Discrepancies:

- Remove invalid values (such as '31' or '32' of the crash classification codes: {'01', '02', '03', 04'})
- Convert typos (such as '.3' and '3' to '03') for consistency
- Mapping codes to their description
  - 'CRASH CLASSIFICATION CODE' ---> 'CRASH CLASSIFICATION DESCRIPTION'
  - 'MANNER OF IMPACT CODE' ---> 'MANNER OF IMPACT DESCRIPTION'
  - 'PRIMARY CONTRIBUTING CIRCUMSTANCE CODE' ---> 'PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION'

Missing Values:

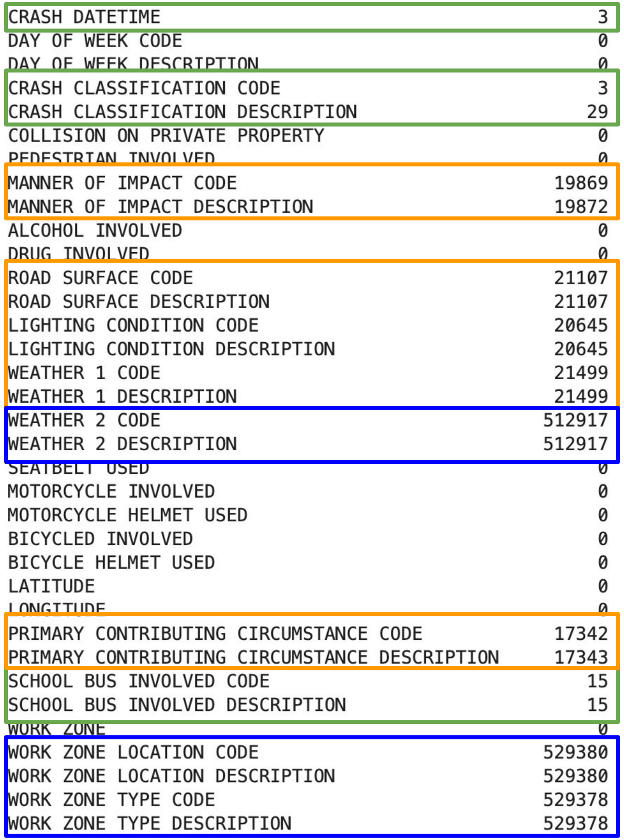
- 0% ---> Do nothing
- < 1% ---> Drop rows with missing values (3)
- < 5% ---> Fill in with most frequent value (4)
- > 5% ---> Drop columns entirely (2)

Data Cleaning Methodology



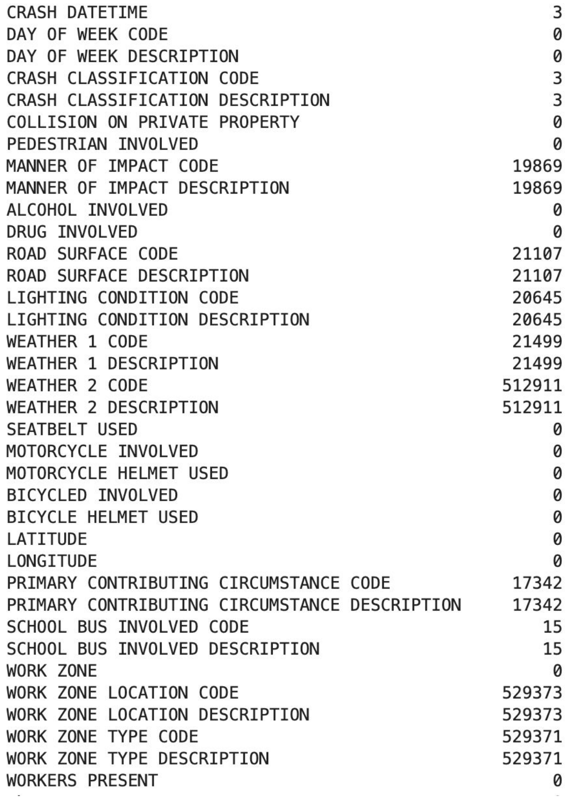
```
CRASH DATETIME                                        3
DAY OF WEEK CODE                                      0
DAY OF WEEK DESCRIPTION                               0
CRASH CLASSIFICATION CODE                             3
CRASH CLASSIFICATION DESCRIPTION                     29
COLLISION ON PRIVATE PROPERTY                         0
PEDESTRIAN INVOLVED                                   0
MANNER OF IMPACT CODE                             19869
MANNER OF IMPACT DESCRIPTION                      19872
ALCOHOL INVOLVED                                      0
DRUG INVOLVED                                         0
ROAD SURFACE CODE                                 21107
ROAD SURFACE DESCRIPTION                          21107
LIGHTING CONDITION CODE                           20645
LIGHTING CONDITION DESCRIPTION                    20645
WEATHER 1 CODE                                    21499
WEATHER 1 DESCRIPTION                             21499
WEATHER 2 CODE                                   512917
WEATHER 2 DESCRIPTION                            512917
SEATBELT USED                                         0
MOTORCYCLE INVOLVED                                   0
MOTORCYCLE HELMET USED                                0
BICYCLED INVOLVED                                     0
BICYCLE HELMET USED                                   0
LATITUDE                                              0
LONGITUDE                                             0
PRIMARY CONTRIBUTING CIRCUMSTANCE CODE            17342
PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION     17343
SCHOOL BUS INVOLVED CODE                             15
SCHOOL BUS INVOLVED DESCRIPTION                      15
WORK ZONE                                             0
WORK ZONE LOCATION CODE                          529380
WORK ZONE LOCATION DESCRIPTION                   529380
WORK ZONE TYPE CODE                              529378
WORK ZONE TYPE DESCRIPTION                       529378
```

Data Cleaning Methodology

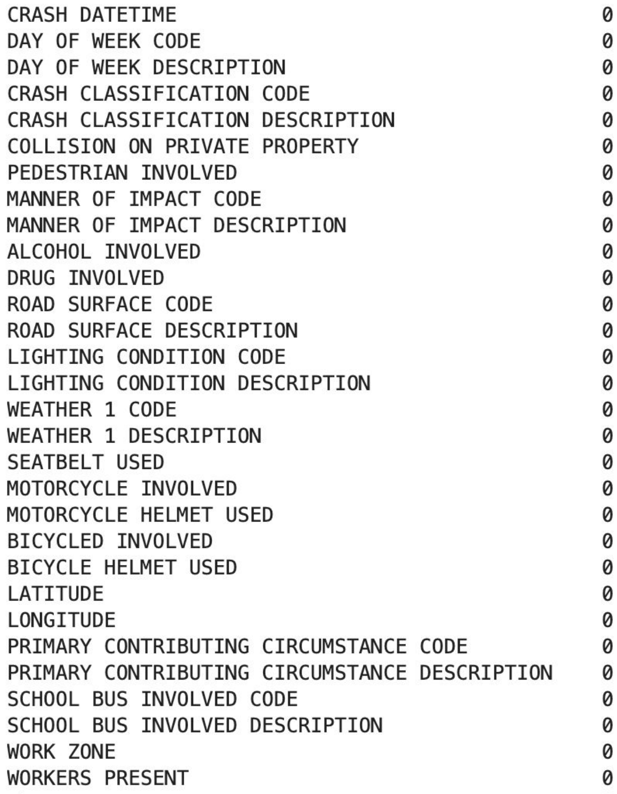


```
CRASH DATETIME                                    3
DAY OF WEEK CODE                                  0
DAY OF WEEK DESCRIPTION                           0
CRASH CLASSIFICATION CODE                         3
CRASH CLASSIFICATION DESCRIPTION                 29
COLLISION ON PRIVATE PROPERTY                     0
PEDESTRIAN INVOLVED                               0
MANNER OF IMPACT CODE                         19869
MANNER OF IMPACT DESCRIPTION                  19872
ALCOHOL INVOLVED                                  0
DRUG INVOLVED                                     0
ROAD SURFACE CODE                             21107
ROAD SURFACE DESCRIPTION                      21107
LIGHTING CONDITION CODE                       20645
LIGHTING CONDITION DESCRIPTION                20645
WEATHER 1 CODE                                21499
WEATHER 1 DESCRIPTION                         21499
WEATHER 2 CODE                               512917
WEATHER 2 DESCRIPTION                        512917
SEATBELT USED                                     0
MOTORCYCLE INVOLVED                               0
MOTORCYCLE HELMET USED                            0
BICYCLED INVOLVED                                 0
BICYCLE HELMET USED                               0
LATITUDE                                          0
LONGITUDE                                         0
PRIMARY CONTRIBUTING CIRCUMSTANCE CODE        17342
PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION 17343
SCHOOL BUS INVOLVED CODE                         15
SCHOOL BUS INVOLVED DESCRIPTION                  15
WORK ZONE                                         0
WORK ZONE LOCATION CODE                      529380
WORK ZONE LOCATION DESCRIPTION               529380
WORK ZONE TYPE CODE                          529378
WORK ZONE TYPE DESCRIPTION                   529378
```

Data Cleaning Methodology



```
CRASH DATETIME                                    3
DAY OF WEEK CODE                                  0
DAY OF WEEK DESCRIPTION                           0
CRASH CLASSIFICATION CODE                         3
CRASH CLASSIFICATION DESCRIPTION                  3
COLLISION ON PRIVATE PROPERTY                     0
PEDESTRIAN INVOLVED                               0
MANNER OF IMPACT CODE                         19869
MANNER OF IMPACT DESCRIPTION                  19869
ALCOHOL INVOLVED                                  0
DRUG INVOLVED                                     0
ROAD SURFACE CODE                             21107
ROAD SURFACE DESCRIPTION                      21107
LIGHTING CONDITION CODE                       20645
LIGHTING CONDITION DESCRIPTION                20645
WEATHER 1 CODE                                21499
WEATHER 1 DESCRIPTION                         21499
WEATHER 2 CODE                               512911
WEATHER 2 DESCRIPTION                        512911
SEATBELT USED                                     0
MOTORCYCLE INVOLVED                               0
MOTORCYCLE HELMET USED                            0
BICYCLED INVOLVED                                 0
BICYCLE HELMET USED                               0
LATITUDE                                          0
LONGITUDE                                         0
PRIMARY CONTRIBUTING CIRCUMSTANCE CODE        17342
PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION 17342
SCHOOL BUS INVOLVED CODE                         15
SCHOOL BUS INVOLVED DESCRIPTION                  15
WORK ZONE                                         0
WORK ZONE LOCATION CODE                      529373
WORK ZONE LOCATION DESCRIPTION               529373
WORK ZONE TYPE CODE                          529371
WORK ZONE TYPE DESCRIPTION                   529371
WORKERS PRESENT                                   0
```



```
CRASH DATETIME                                    0
DAY OF WEEK CODE                                  0
DAY OF WEEK DESCRIPTION                           0
CRASH CLASSIFICATION CODE                         0
CRASH CLASSIFICATION DESCRIPTION                  0
COLLISION ON PRIVATE PROPERTY                     0
PEDESTRIAN INVOLVED                               0
MANNER OF IMPACT CODE                             0
MANNER OF IMPACT DESCRIPTION                      0
ALCOHOL INVOLVED                                  0
DRUG INVOLVED                                     0
ROAD SURFACE CODE                                 0
ROAD SURFACE DESCRIPTION                          0
LIGHTING CONDITION CODE                           0
LIGHTING CONDITION DESCRIPTION                    0
WEATHER 1 CODE                                    0
WEATHER 1 DESCRIPTION                             0
SEATBELT USED                                     0
MOTORCYCLE INVOLVED                               0
MOTORCYCLE HELMET USED                            0
BICYCLED INVOLVED                                 0
BICYCLE HELMET USED                               0
LATITUDE                                          0
LONGITUDE                                         0
PRIMARY CONTRIBUTING CIRCUMSTANCE CODE            0
PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION     0
SCHOOL BUS INVOLVED CODE                          0
SCHOOL BUS INVOLVED DESCRIPTION                   0
WORK ZONE                                         0
WORKERS PRESENT                                   0
```

EDA

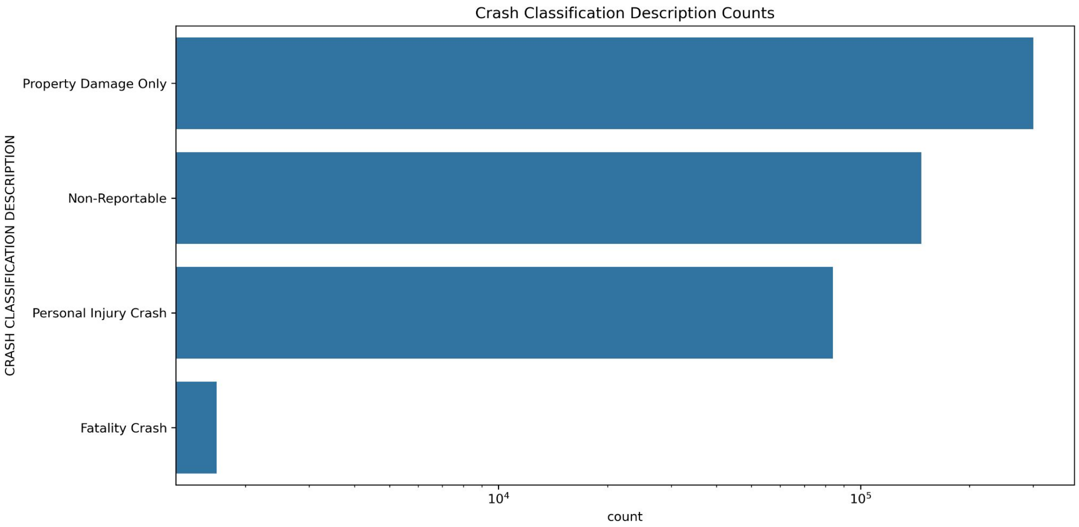


Fig. 1. Distribution of crash classifications. The horizontal bar plot shows the total number of crashes for each classification, with the x-axis scaled logarithmically.

Note: the classes are highly imbalanced, with “Fatality Crash” far less than the other three categories.

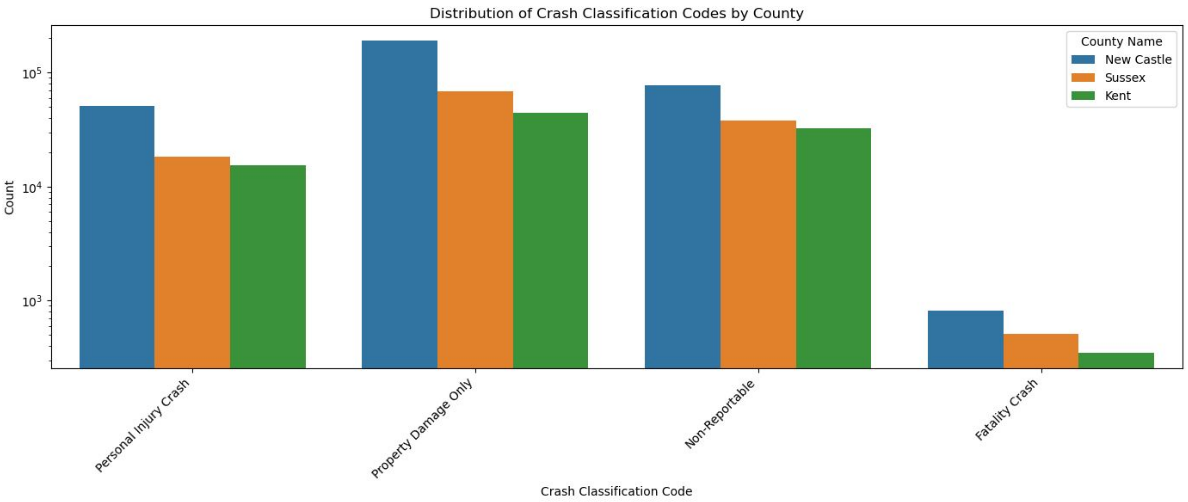
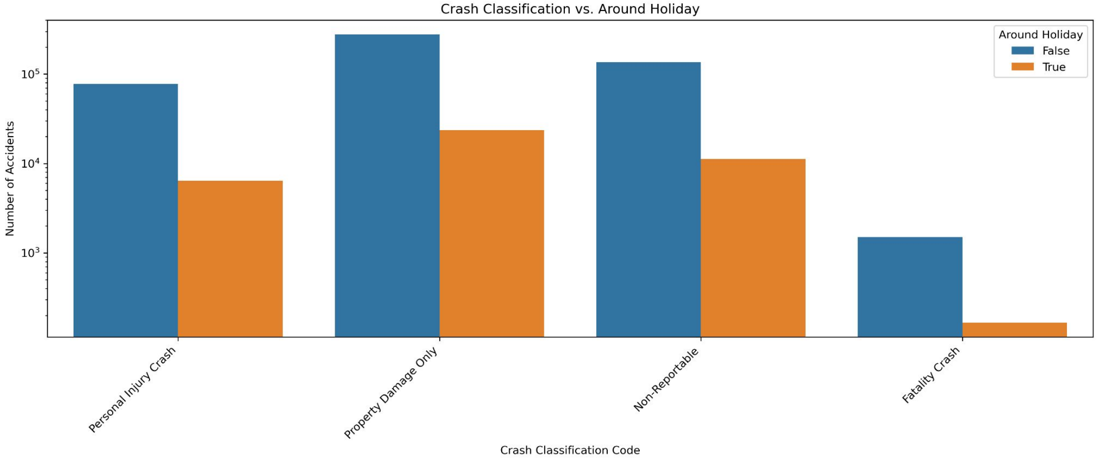
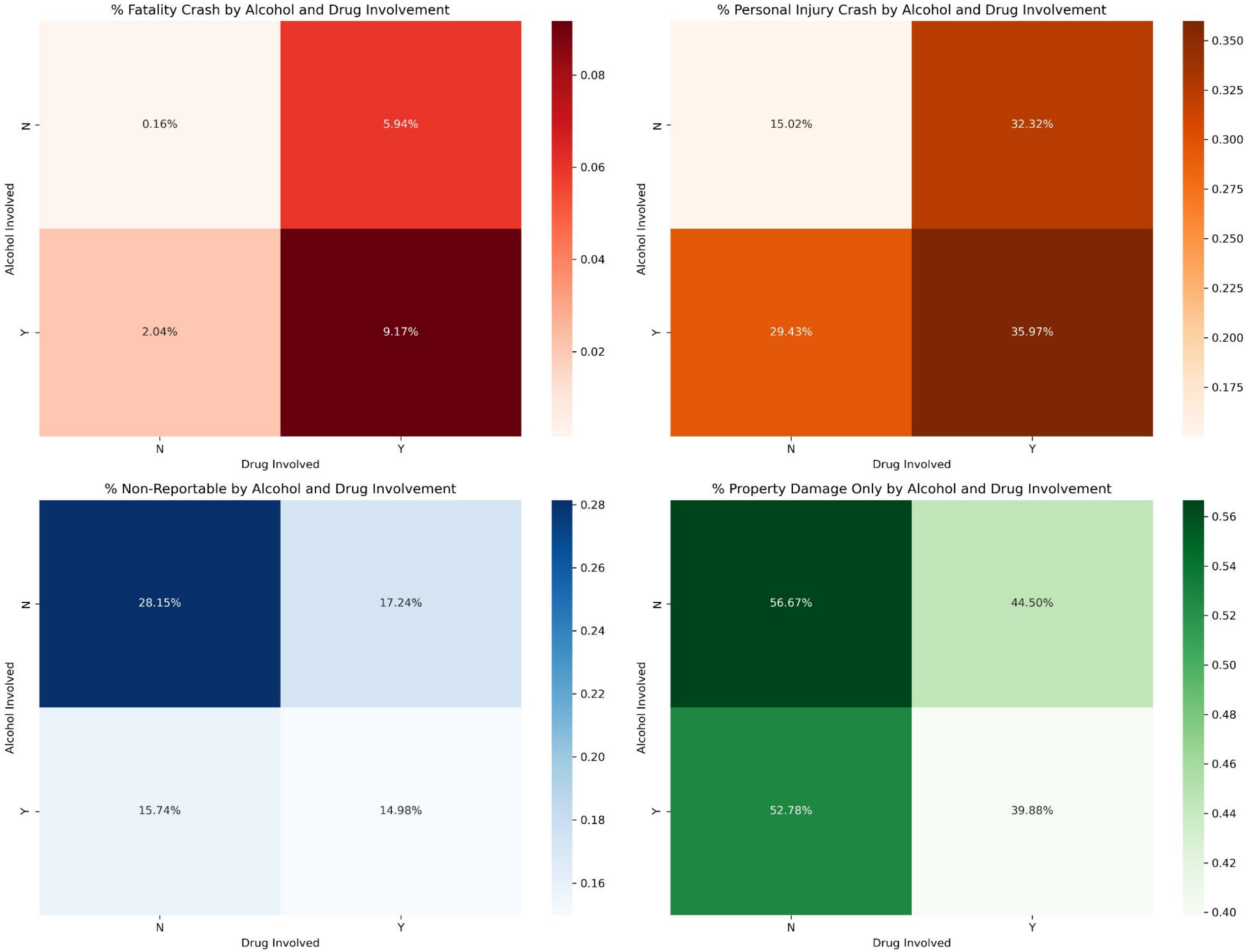


Fig. 2. (Left) Distribution of crash classification codes by county with the y-axis scaled logarithmically.

Fig. 3. (Right) Distribution of crash classification codes by whether the crash occurred around a holiday, with the y-axis scaled logarithmically. Holiday time frames are defined as a 5-day window centered around a significant US holiday.



Note: both plots reflect the class imbalance present in the data.



% Fatality Crash by Alcohol and Drug Involvement

% Personal Injury Crash by Alcohol and Drug Involvement

% Non-Reportable by Alcohol and Drug Involvement

% Property Damage Only by Alcohol and Drug Involvement

Fig. 4. Heatmap of alcohol and drug by crash classification. The heatmaps show the percentage of crashes within each classification that involved combinations of alcohol and drug involvement.

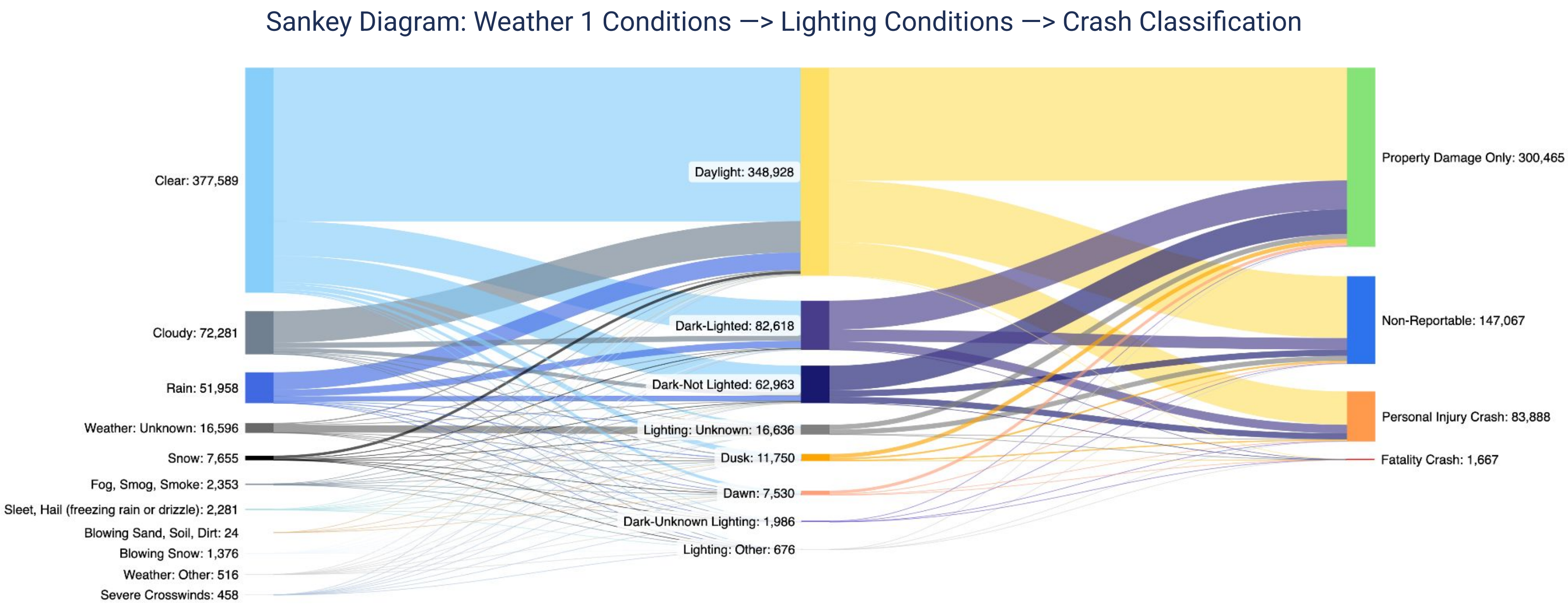


Fig. 5. Sankey diagram of crash conditions. Flow paths show how crashes progress from initial weather conditions (1) to lighting conditions, and finally to crash classification.

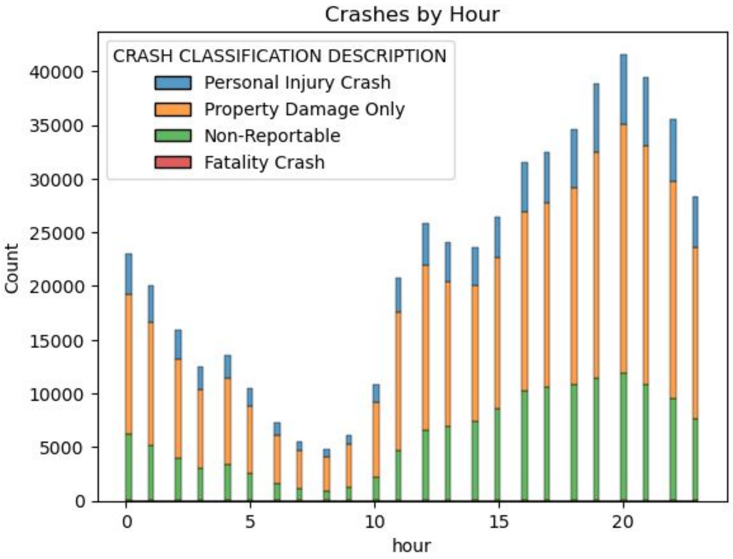


Fig. 6. Univariate distribution of crash counts over a 24-hour period, stratified by crash class classification.

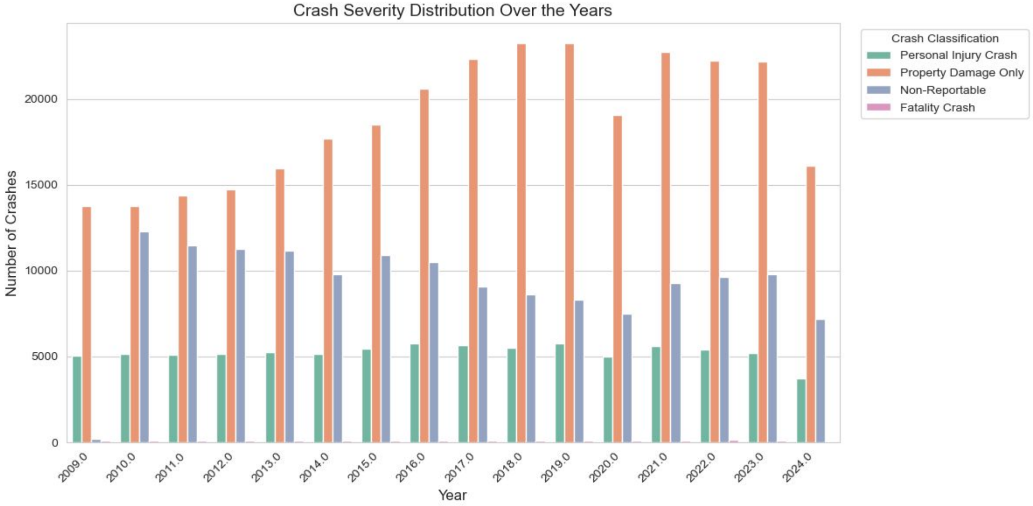


Fig. 7. Univariate distribution of crash counts over a 16-year period, stratified by crash class classification.

Feature Engineering

Techniques:

- One-Hot Encoding
  - Applied to non-binary converted columns; categorical data (e.g `'COLLISION ON PRIVATE PROPERTY', 'COUNTY NAME'`)
- Binary Mapping
  - Converting **Yes/No** features to 1/0 using mapping `binary_mapping = {'Y': 1, 'YES': 1, '1': 1, 'N': 0, 'NO': 0, '0': 0}` e.g (`'ALCOHOL INVOLVED', 'SEATBELT USED'`)
- Feature Selection
  - Used Cramer's V to detect and drop redundant nominal categorical features
  - Used Mutual Information
  - Some features are extremely imbalance resulting in low MI, however from proportion distribution, some of them does show distinctive impact on target so we kept for unique influence

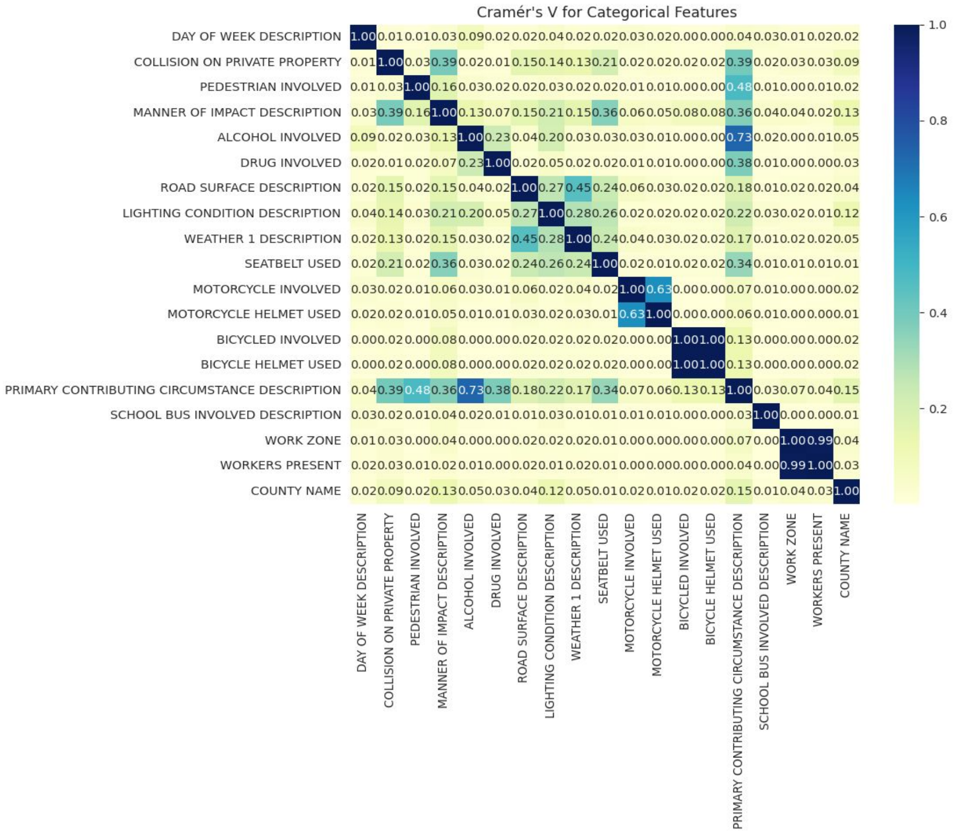


Fig. 8. Correlation matrix of 23 features out of 37 total features.

Moderate Redundancy:

ROAD SURFACE DESCRIPTION, WEATHER 1 DESCRIPTION, LIGHTING CONDITION DESCRIPTION

—> Suggest overlapping information e.g., poor weather -> wet roads -> poor visibility.

Strong Dependency:

MOTORCYCLE INVOLVED ↔ MOTORCYCLE HELMET USED (V = 0.63)

BICYCLE INVOLVED ↔ BICYCLE HELMET USED (V = 1.00)

WORK ZONE ↔ WORKERS PRESENT (V ≈ 0.99)

—>These features are nearly duplicates; we retain only one from each pair.

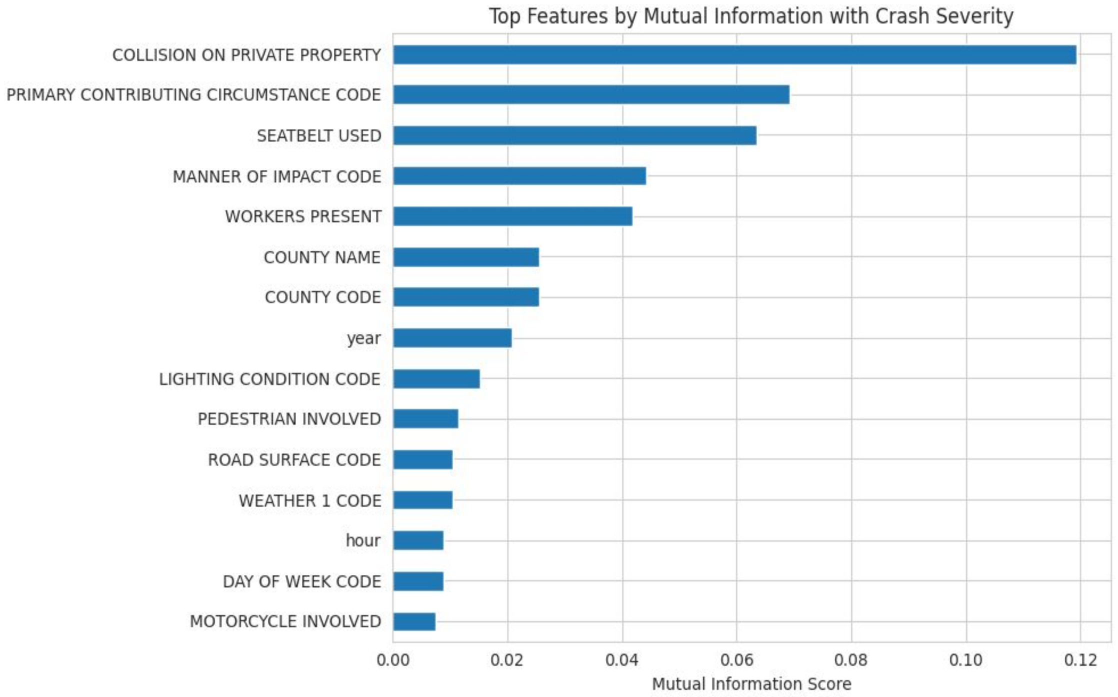
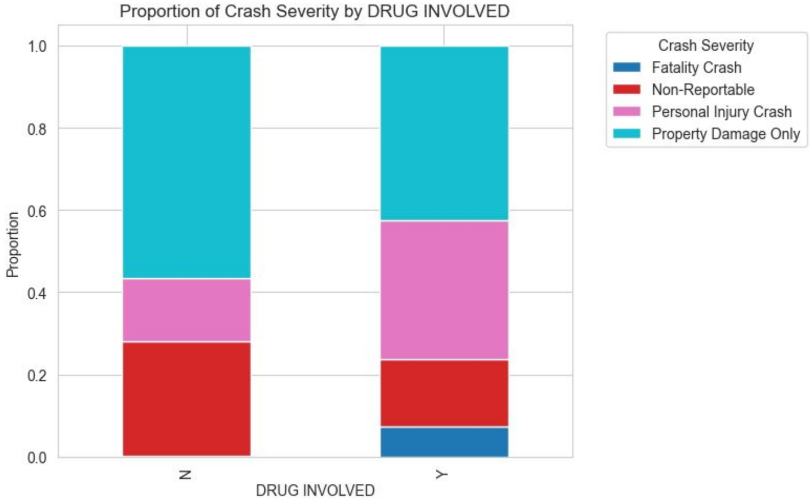


Fig. 9. Top feature distribution by mutual information.

WHY?

- MI measures how much information a feature provides about the target
- It can capture both categorical and numerical relationships (unlike correlation, which only sees linear relationships)



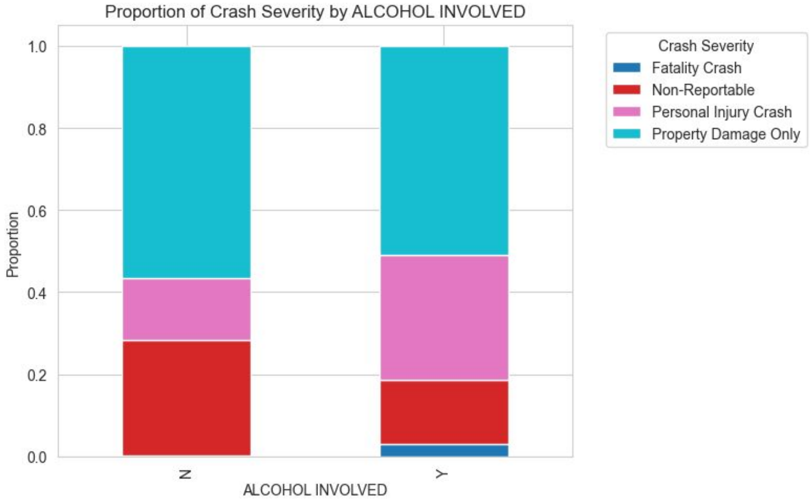


Fig. 10. Distribution of proportions of crash classification, categorized by drug involvement.

Fig. 11. Distribution of proportions of crash classification, categorized by alcohol involvement.

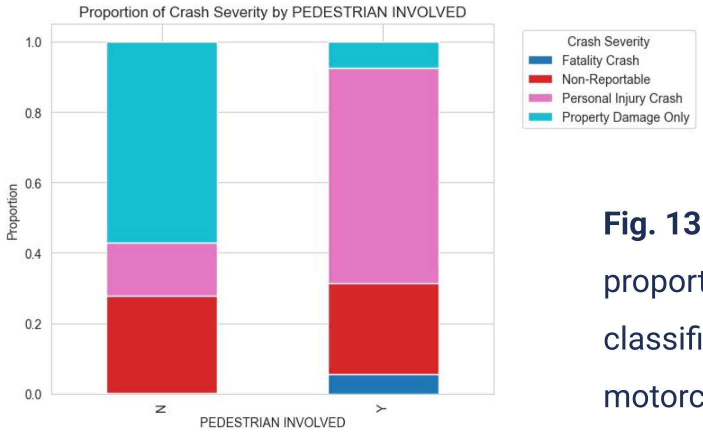


Fig. 13. (Below) Distribution of proportions of crash classification, categorized by motorcycle involvement.

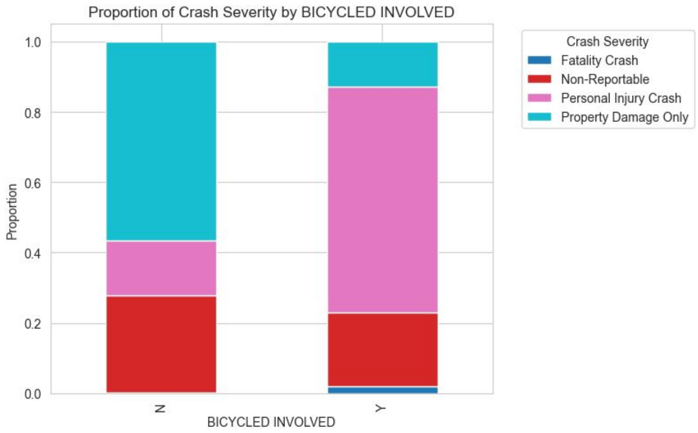


Fig. 12. (Above) Distribution of proportions of crash classification, categorized by pedestrian involvement.

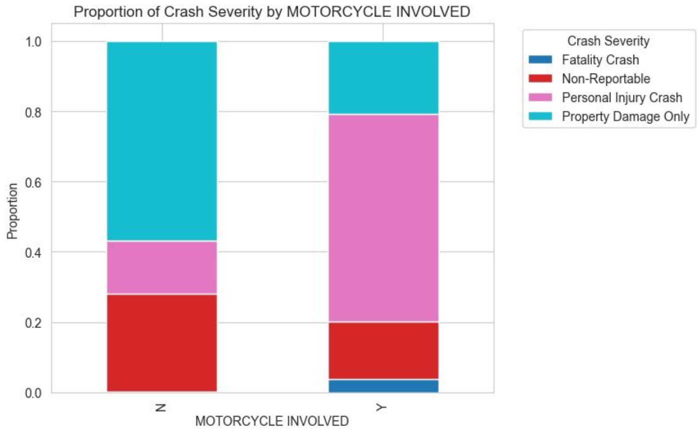


Fig. 14. (Above) Distribution of proportions of crash classification, categorized by bicycle involvement.

Note: all three graphs' involvement share heavy weight towards personal injury crash

Data Diagnostics + Transformation

PCA:

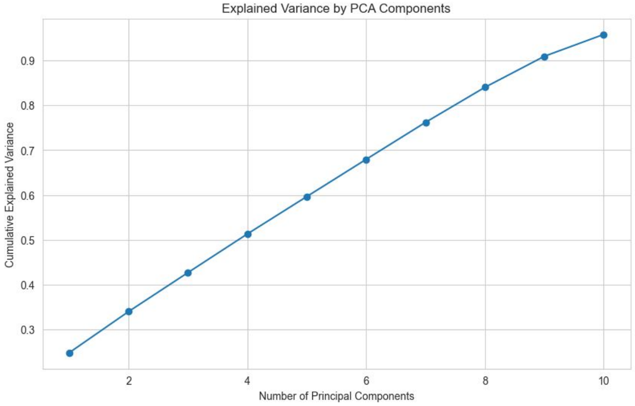
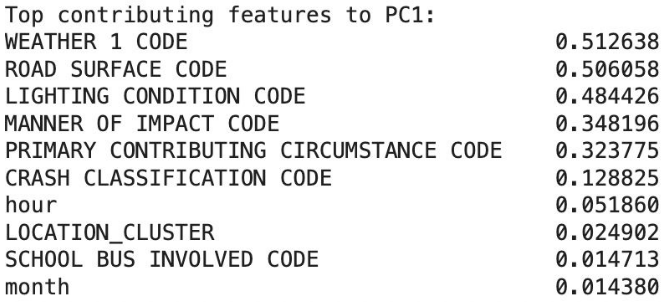


Fig. 15. (Left) Line plot showing the explained variance ratio by principal component.



```
Top contributing features to PC1:
WEATHER 1 CODE                            0.512638
ROAD SURFACE CODE                         0.506058
LIGHTING CONDITION CODE                   0.484426
MANNER OF IMPACT CODE                     0.348196
PRIMARY CONTRIBUTING CIRCUMSTANCE CODE    0.323775
CRASH CLASSIFICATION CODE                 0.128825
hour                                      0.051860
LOCATION_CLUSTER                          0.024902
SCHOOL BUS INVOLVED CODE                  0.014713
month                                     0.014380
```

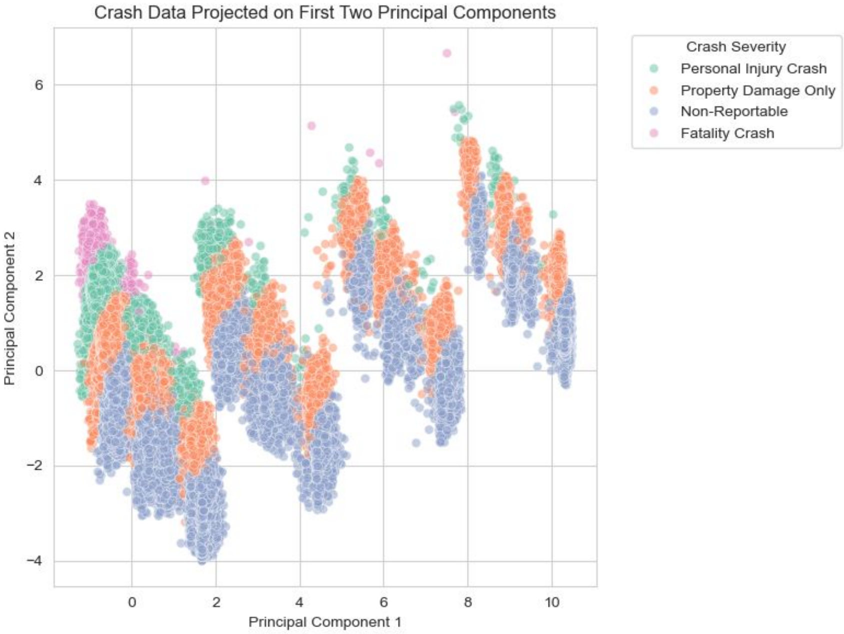
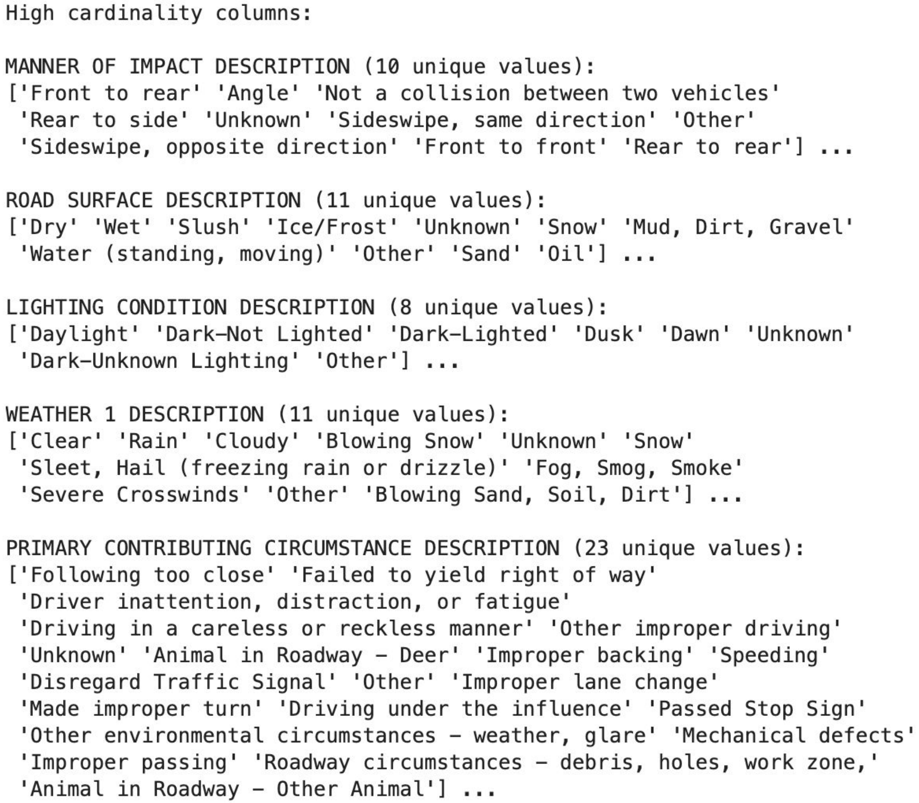


Fig. 16. Crash data projected onto the first two principal components. Each point represents a crash instance.

Note: less variability amongst the fatality crash group

K-Means Clustering



```
High cardinality columns:

MANNER OF IMPACT DESCRIPTION (10 unique values):
['Front to rear' 'Angle' 'Not a collision between two vehicles'
 'Rear to side' 'Unknown' 'Sideswipe, same direction' 'Other'
 'Sideswipe, opposite direction' 'Front to front' 'Rear to rear'] ...

ROAD SURFACE DESCRIPTION (11 unique values):
['Dry' 'Wet' 'Slush' 'Ice/Frost' 'Unknown' 'Snow' 'Mud, Dirt, Gravel'
 'Water (standing, moving)' 'Other' 'Sand' 'Oil'] ...

LIGHTING CONDITION DESCRIPTION (8 unique values):
['Daylight' 'Dark-Not Lighted' 'Dark-Lighted' 'Dusk' 'Dawn' 'Unknown'
 'Dark-Unknown Lighting' 'Other'] ...

WEATHER 1 DESCRIPTION (11 unique values):
['Clear' 'Rain' 'Cloudy' 'Blowing Snow' 'Unknown' 'Snow'
 'Sleet, Hail (freezing rain or drizzle)' 'Fog, Smog, Smoke'
 'Severe Crosswinds' 'Other' 'Blowing Sand, Soil, Dirt'] ...

PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION (23 unique values):
['Following too close' 'Failed to yield right of way'
 'Driver inattention, distraction, or fatigue'
 'Driving in a careless or reckless manner' 'Other improper driving'
 'Unknown' 'Animal in Roadway - Deer' 'Improper backing' 'Speeding'
 'Disregard Traffic Signal' 'Other' 'Improper lane change'
 'Made improper turn' 'Driving under the influence' 'Passed Stop Sign'
 'Other environmental circumstances - weather, glare' 'Mechanical defects'
 'Improper passing' 'Roadway circumstances - debris, holes, work zone,'
 'Animal in Roadway - Other Animal'] ...
```

Dry

Wet, Slush, Ice/Frost, Snow

Mud, Dirt, Gravel, Sand

Bright lighting

Some lighting

Low lighting

Driver negligence

Breaking traffic rules

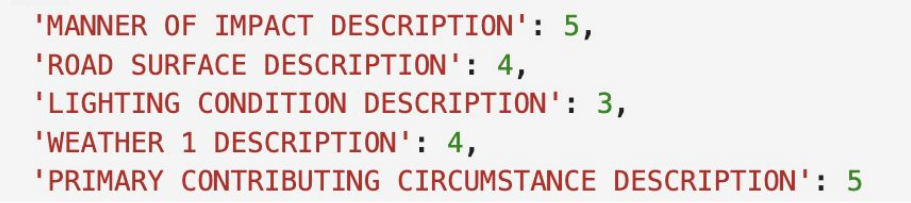
Under the influence

Outside factors

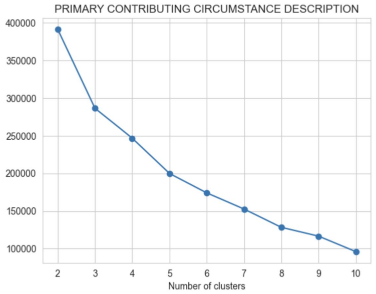
Speeding

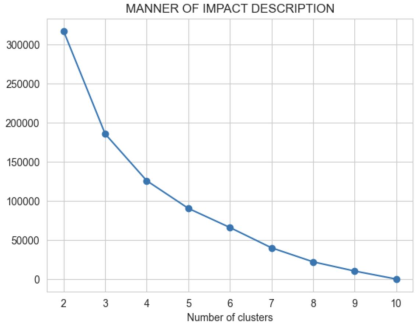
K-Means Clustering

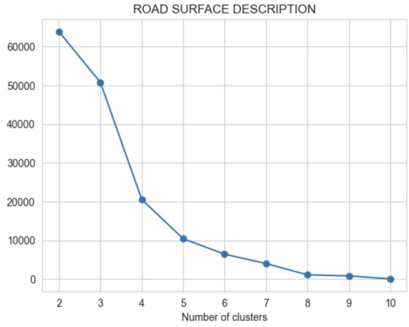
Selecting **K** using elbow method:

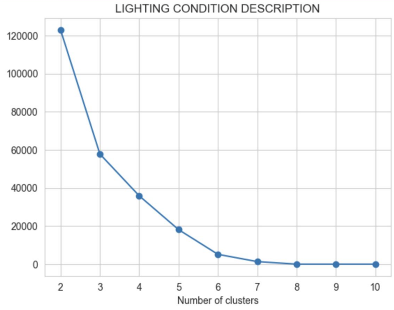


```
'MANNER OF IMPACT DESCRIPTION': 5,
'ROAD SURFACE DESCRIPTION': 4,
'LIGHTING CONDITION DESCRIPTION': 3,
'WEATHER 1 DESCRIPTION': 4,
'PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION': 5
```


PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION


MANNER OF IMPACT DESCRIPTION


ROAD SURFACE DESCRIPTION


LIGHTING CONDITION DESCRIPTION

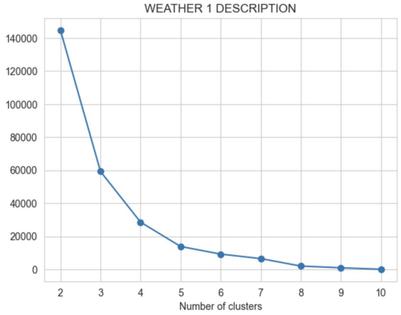

WEATHER 1 DESCRIPTION

Fig. 17-21. Elbow plots for columns with large number of unique values (num. of clusters vs within cluster sum of squares).

K-Means Clustering

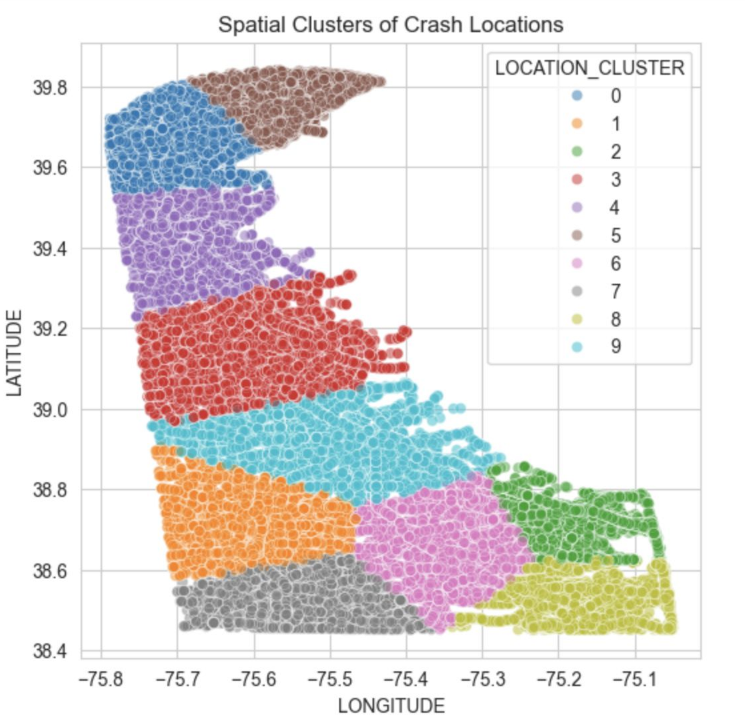
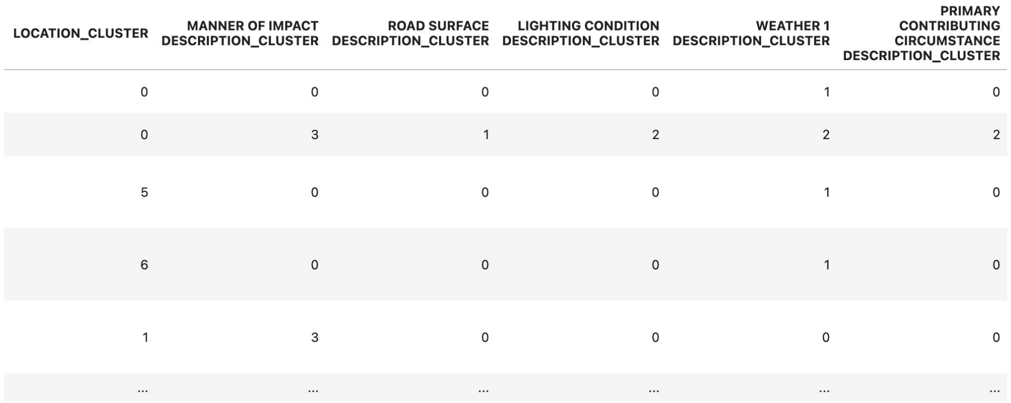


Fig. 22. Latitude/longitude clustering of crash data in Delaware.

We created 6 new features using clustering:

- 1 location clustering based on latitude/longitude data
- 5 condensed high-cardinality columns from the original data



LOCATION_CLUSTER	MANNER OF IMPACT DESCRIPTION_CLUSTER	ROAD SURFACE DESCRIPTION_CLUSTER	LIGHTING CONDITION DESCRIPTION_CLUSTER	WEATHER 1 DESCRIPTION_CLUSTER	PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION_CLUSTER
0	0	0	0	1	0
0	3	1	2	2	2
5	0	0	0	1	0
6	0	0	0	1	0
1	3	0	0	0	0
...	...	...	...	...	...

Addressing Class Imbalance

- Undersampling
- Oversampling (SMOTE)
- Class weights

3-Fold Cross-Validation (due to hardware constraints)

	Accuracy	Recall	F1-Score
Baseline	0.6324	0.3927	0.4096
Class Weights	0.6237	0.3847	0.4002
SMOTE	0.6538	0.4170	0.4379
Undersampling	0.4462	0.5118	0.3460

Classification Modeling

- Due to runtime efficiency, we will be using the baseline model for:
  - Logistic regression
  - Random forest
  - K-nearest neighbor
  - XGBoost
- Once a final model is selected, we will use best subset variables as it provides equal performance to utilizing all variables, plus:
  - Makes analysis easier + more simple to interpret
  - Faster prediction times
  - Less noise and highlights important features

XGBoost

Accuracy: 0.6784

	Precision	Recall	F1-Score	Support
Non-Reportable	0.78	0.48	0.59	29588
Property Damage Only	0.66	0.93	0.77	60518
Personal Injury Crash	0.65	0.15	0.24	16877
Fatality Crash	0.55	0.16	0.25	335
Accuracy			0.68	107318
Macro Avg	0.66	0.43	0.46	107318
Weighted Avg	0.69	0.68	0.64	107318

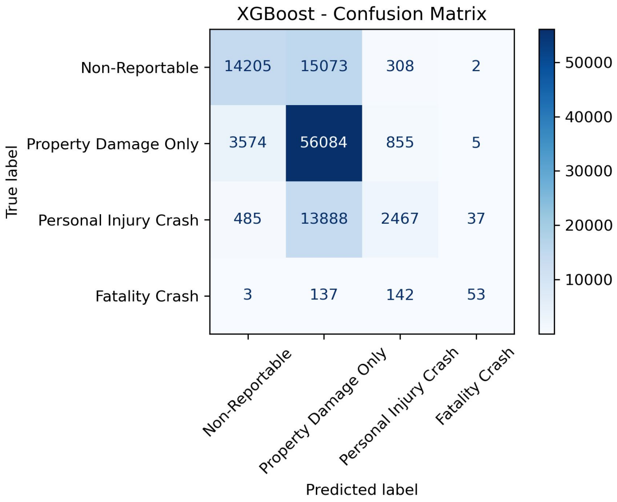
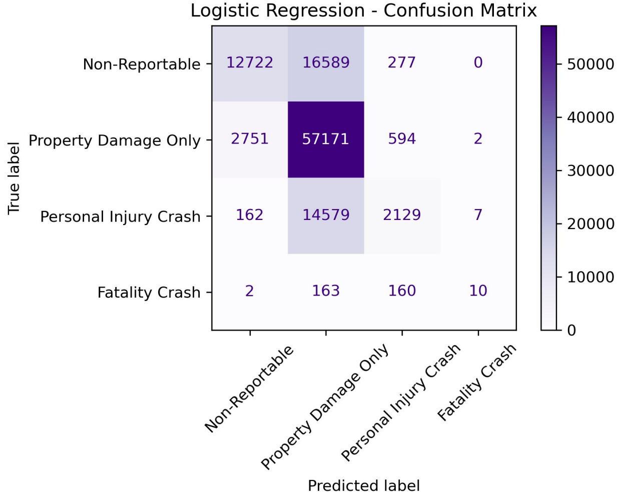


Fig. 23. Confusion matrix showing test set predictions by the XGBoost classifier trained without resampling (baseline model).

Logistic Regression

Accuracy: 0.6712

	Precision	Recall	F1-Score	Support
Non-Reportable	0.81	0.43	0.56	29588
Property Damage Only	0.65	0.94	0.77	60518
Personal Injury Crash	0.67	0.13	0.21	16877
Fatality Crash	0.53	0.03	0.06	335
Accuracy			0.67	107318
Macro Avg	0.66	0.38	0.40	107318
Weighted Avg	0.70	0.67	0.62	107318



True label \ Predicted label	Non-Reportable	Property Damage Only	Personal Injury Crash	Fatality Crash
Non-Reportable	12722	16589	277	0
Property Damage Only	2751	57171	594	2
Personal Injury Crash	162	14579	2129	7
Fatality Crash	2	163	160	10

Fig. 24. Confusion matrix showing test set predictions by the Logistic Regression classifier trained without resampling (baseline model).

Random Forest

Accuracy: 0.6175

	Precision	Recall	F1-Score	Support
Non-Reportable	0.61	0.53	0.57	29588
Property Damage Only	0.66	0.76	0.71	60518
Personal Injury Crash	0.38	0.27	0.31	16877
Fatality Crash	0.36	0.14	0.20	335
Accuracy			0.62	107318
Macro Avg	0.50	0.42	0.45	107318
Weighted Avg	0.60	0.62	0.61	107318

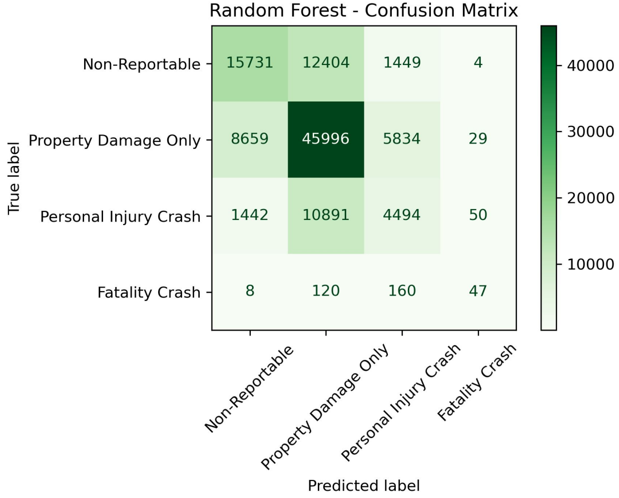


Fig. 25. Confusion matrix showing test set predictions by the Random Forest classifier trained without resampling (baseline model).

KNN

Accuracy: 0.6166

	Precision	Recall	F1-Score	Support
Non-Reportable	0.57	0.53	0.55	29588
Property Damage Only	0.65	0.78	0.71	60518
Personal Injury Crash	0.42	0.19	0.26	16877
Fatality Crash	0.45	0.04	0.08	335
Accuracy			0.62	107318
Macro Avg	0.53	0.39	0.40	107318
Weighted Avg	0.59	0.62	0.59	107318

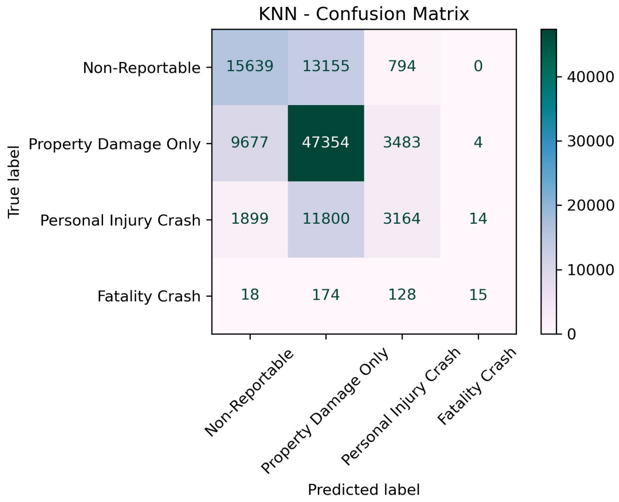


Fig. 26. Confusion matrix showing test set predictions by the K-Nearest Neighbors classifier trained without resampling (baseline model).

Model Selection

Model Performance Summary (Accuracy)

	Accuracy
XGBoost	0.678442
Logistic Regression	0.671201
Random Forest	0.617492
KNN	0.616597

- Out of the 4 models, XGBoost has the highest accuracy
  - We will select XGBoost using a baseline model

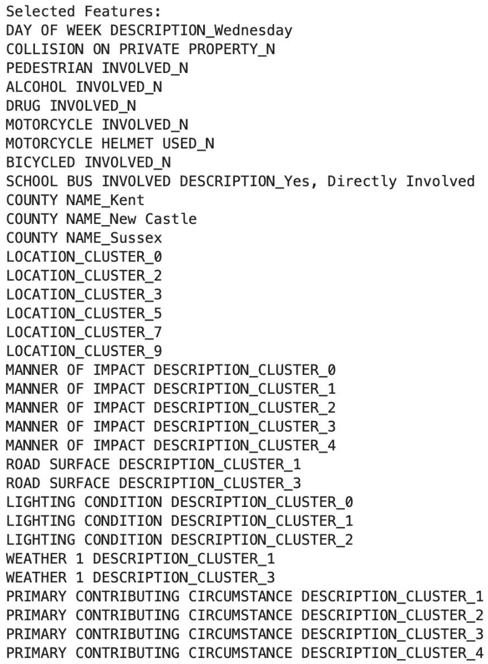
Main Results

- **SMOTE** led to the best results overall, but long runtime and memory constraints made it an unfeasible method for big-data usage
    - Instead we went with the second-best method, the **baseline model** (no resampling)

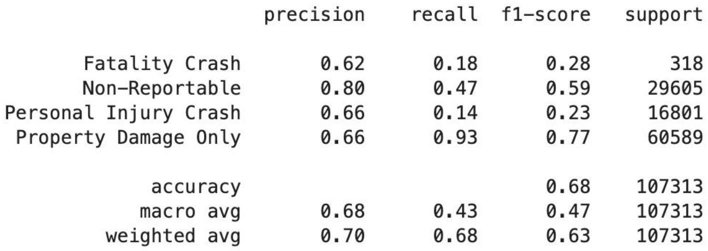
- **XGBoost** obtained the highest prediction accuracy, followed by **Logistic Regression**
    - **Random Forest** and **KNN** had a lower performance in accuracy in comparison
    - After choosing **XGBoost**, we optimized the model to meet our goals of prioritizing recall, in accordance to fatality crashes and personal injuries
        - Best subset columns were used to obtain a balanced recall

Final Model Selection

Our best subset selection minimized on multi-class log loss chose 34 out of 131 one-hot encoded variables (originally 25):



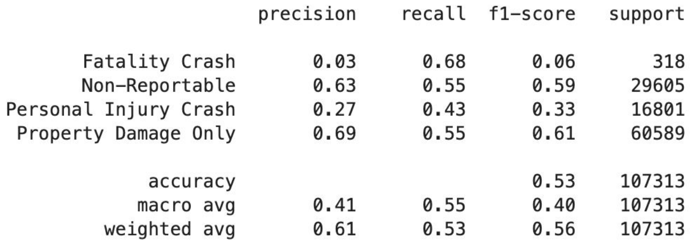
```
Selected Features:
DAY OF WEEK DESCRIPTION_Wednesday
COLLISION ON PRIVATE PROPERTY_N
PEDESTRIAN INVOLVED_N
ALCOHOL INVOLVED_N
DRUG INVOLVED_N
MOTORCYCLE INVOLVED_N
MOTORCYCLE HELMET USED_N
BICYCLED INVOLVED_N
SCHOOL BUS INVOLVED DESCRIPTION_Yes, Directly Involved
COUNTY NAME_Kent
COUNTY NAME_New Castle
COUNTY NAME_Sussex
LOCATION_CLUSTER_0
LOCATION_CLUSTER_2
LOCATION_CLUSTER_3
LOCATION_CLUSTER_5
LOCATION_CLUSTER_7
LOCATION_CLUSTER_9
MANNER OF IMPACT DESCRIPTION_CLUSTER_0
MANNER OF IMPACT DESCRIPTION_CLUSTER_1
MANNER OF IMPACT DESCRIPTION_CLUSTER_2
MANNER OF IMPACT DESCRIPTION_CLUSTER_3
MANNER OF IMPACT DESCRIPTION_CLUSTER_4
ROAD SURFACE DESCRIPTION_CLUSTER_1
ROAD SURFACE DESCRIPTION_CLUSTER_3
LIGHTING CONDITION DESCRIPTION_CLUSTER_0
LIGHTING CONDITION DESCRIPTION_CLUSTER_1
LIGHTING CONDITION DESCRIPTION_CLUSTER_2
WEATHER 1 DESCRIPTION_CLUSTER_1
WEATHER 1 DESCRIPTION_CLUSTER_3
PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION_CLUSTER_1
PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION_CLUSTER_2
PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION_CLUSTER_3
PRIMARY CONTRIBUTING CIRCUMSTANCE DESCRIPTION_CLUSTER_4
```



```
                       precision    recall  f1-score   support

       Fatality Crash       0.62      0.18      0.28       318
       Non-Reportable       0.80      0.47      0.59     29605
 Personal Injury Crash       0.66      0.14      0.23     16801
  Property Damage Only       0.66      0.93      0.77     60589

             accuracy                           0.68    107313
            macro avg       0.68      0.43      0.47    107313
         weighted avg       0.70      0.68      0.63    107313
```



```
                       precision    recall  f1-score   support

       Fatality Crash       0.03      0.68      0.06       318
       Non-Reportable       0.63      0.55      0.59     29605
 Personal Injury Crash       0.27      0.43      0.33     16801
  Property Damage Only       0.69      0.55      0.61     60589

             accuracy                           0.53    107313
            macro avg       0.41      0.55      0.40    107313
         weighted avg       0.61      0.53      0.56    107313
```

Discussion of Findings

- We adopted the **baseline model** over **SMOTE**, despite **SMOTE** being an overall better performer due to long runtimes and memory issues
    - Opting instead for the **baseline model** trades slight performance improvement for better efficiency + preferable scaling
- We prioritize recall due to importance of identifying severe and fatal crashes
- Although best subset didn't improve accuracy, it helped stabilized recall, therefore we optimized **XGBoost** to prioritize recall

Interesting Findings in Context:

—> "Bicycle involved" was present, yet "bicycle helmet used" wasn't, however "motorcycle involved" and "helmet" were

The bicycles variables also had a 1.0 correlation; in our data, all bikers who got in an accident wore their helmet

Not all motorcyclist wear them and "motorcycle helmet" is a separate and significant variable in class severity; WEAR HELMETS

—> Not all location clusters were used, some regions of Delaware have more unsafe traffic conditions than others

—> Wednesdays are significant for crashes, only two weather clusters and two road conditions were significant

—> Substance abuse was expected, however school busses are rare and remained in the final model; people drive safer around busses

Conclusions

Limitation:

- Class Imbalance in Target Variable: the target is heavily imbalanced (e.g., very few “Fatality Crash” cases) so even with resampling, there are still risks like oversampling which can lead to overfitting or undersampling may discard potentially valuable data
- Missing or Incomplete Data: Several features (e.g ‘Weather 2’, ‘Work zone location’) were dropped because they have >90% missing values that could cause potential loss of information
- Traffic patterns and road conditions may change in the future so it needs to be updated if used on future data

Improvement:

❖ Expanding the scope of the data to other states or to multi-year data

❖ Incorporate with real-time applications for richer data

❖ Apply a combination of sampling techniques to effectively handle class imbalance