**Factors Impacting Motor Vehicle Crash Severity**

Main Report

Author’s Note: Discussion of the data set and the methods will be performed here. Analysis of the results is contained within the Jupyter notebook to keep it close to the code and results. A copy of the notebook is included at the end of this document.

**Data**

The Allegheny County Crash Data data set is available from data.gov and is provided by the Allegheny County/ City of Pittsburg/ Western PA Regional Data Center. The description from <https://catalog.data.gov/dataset/allegheny-county-crash-data> reads:

*Contains locations and information about every crash incident reported to the police in Allegheny County from 2004 to 2017. Fields include injury severity, fatalities, information about the vehicles involved, location information, and factors that may have contributed to the crash.*

The data set contains 170,358 crash incident records and 189 features. 68 features were removed prior to any analysis for being either redundant or not relevant to the target variable. List 1 in the appendix lists the features removed and why.

The target variable for this analysis was the feature FATAL\_OR\_MAJ\_INJ. A value of 1 was assigned if the incident had at least one major injury or fatality. Another feature captured a more specific severity level (MAX\_SEVERITY\_LEVEL). Two of the possible values were 8 – Injury/Unknown Severity and 9 – Unknown. Because one the tasks was to attempt to predict FATAL\_OR\_MAJ\_INJ, any records with 8 or 9 for MAX\_SEVERITY\_LEVEL were removed.

Next, I proceeded to determine if there were any missing values in the remaining data and handle them.

APPENDIX

**List 1 – Features Removed**

DISTRICT: only one value

CRASH\_COUNTY: only one value

POLICE\_AGENCY: not relevant

TIME\_OF\_DAY: too specific, prefer hour

SCH\_BUS\_IND: redundant, missing data

SCH\_ZONE\_IND: redundant, missing data

TOTAL\_UNITS: undefined meaning

FATAL\_COUNT: redundant

INJURY\_COUNT: redundant

MAJ\_INJ\_COUNT: redundant

MOD\_INJ\_COUNT: redundant

MIN\_INJ\_COUNT: redundant

UNK\_INJ\_DEG\_COUNT: excluded

UNK\_INJ\_PER\_COUNT: excluded

UNB\_DEATH\_COUNT: not relevant

UNB\_MAJ\_INJ\_COUNT: not relevant

BELTED\_DEATH\_COUNT: not relevant

BELTED\_MAJ\_INJ\_COUNT: not relevant

MCYCLE\_DEATH\_COUNT: not relevant

MCYCLE\_MAJ\_INJ\_COUNT: not relevant

BICYCLE\_DEATH\_COUNT: not relevant

BICYCLE\_MAJ\_INJ\_COUNT: not relevant

PED\_DEATH\_COUNT: not relevant

PED\_MAJ\_INJ\_COUNT: not relevant

MAX\_SEVERITY\_LEVEL: redundant

LATITUDE: not relevant

LONGITUDE: not relevant

DEC\_LAT: not relevant

DEC\_LONG: not relevant

EST\_HRS\_CLOSED: not relevant

LANE\_CLOSED: not relevant (interpreted as lane closure as a result of the crash)

LN\_CLOSE\_DIR: not relevant

NTFY\_HIWY\_MAINT: not relevant

WZ\_CLOSE\_DETOUR: limited data

WZ\_FLAGGER: limited data

WZ\_LAW\_OFFCR\_IND: limited data

WZ\_LN\_CLOSURE: limited data

WZ\_MOVING: limited data

WZ\_OTHER: limited data

WZ\_SHLDER\_MDN: limited data

FLAG\_CRN: redundant

PROPERTY\_DAMAGE\_ONLY: redundant

INJURY: redundant

FATAL: redundant

NON\_INTERSECTION: redundant

DRIVER\_COUNT\_16YR: redundant

DRIVER\_COUNT\_17YR: redundant

DRIVER\_COUNT\_18YR: redundant

DRIVER\_COUNT\_19YR: redundant

DRIVER\_COUNT\_20YR: redundant

DRIVER\_COUNT\_50\_64YR: redundant

DRIVER\_COUNT\_65\_74YR: redundant

DRIVER\_COUNT\_75PLUS: redundant

VEHICLE\_TOWED: not relevant

MINOR\_INJURY: redundant

MODERATE\_INJURY: redundant

MAJOR\_INJURY: redundant

PSP\_REPORTED: not relevant

ROADWAY\_CRN: redundant

RDWY\_SEQ\_NUM: not relevant

ADJ\_RDWY\_SEQ: not relevant

ACCESS\_CTRL: undefined meaning

ROADWAY\_COUNTY: only one value

RDWY\_ORIENT: not relevant

ROUTE: not relevant (think about this one)

TOT\_INJ\_COUNT: redundant

SCHOOL\_BUS\_UNIT: undefined meaning

LIMIT\_65MPH: redundant and contradictory