

car accidents analysis

Capstone report

1 de abril de 2020

# Context

El estado de NY está enfrentando una problemática de congestionamiento vial debido a que se están haciendo reportes de accidentes categorizados como leves cuando en realidad la severidad de los mismo es mayor a la reportada. Por ende, los usuarios de las vías no conocen cuanto tiempo tardará la vía en habilitarse y esto hace que muchas veces tengan que esperar tiempos muy extensos, pudiendo haber tomado una ruta alterna de haber conocido cuánto tiempo aproximado se tardaría.

# Business question

**What is the stated business question?**

¿Es posible predecir la severidad de un accidente en NY?

# Analysis plan

**What is the analysis goal?**

Diseñar, crear e implementar un modelo que permita predecir la **severidad** de un accidente de tránsito en relación con el **tiempo** que se tarda en habilitar la vía desde el momento en que este ocurre.

# Exploratory Data Analysis

La información disponible para realizar el análisis es la siguiente:

|  |  |
| --- | --- |
| **Atributo** | **Descripción** |
| ID | Identificador único del accidente |
| Source | Fuente de reporte del accidente |
| TMC | Traffic Message Channel el cual provee más detalles del evento. |
| **Severity** | Severidad del accidente. Un número entre 1y4. 1 indica el menor impacto en el tráfico vial en términos del tiempo que tarda habilitar la ruta nuevamente.  **1: 1hrs o menos**  **2: 2hrs o menos**  **3: 3hrs o menos**  **4: more tan 4hrs** |
| Start\_Time | Hora de inicio del accidente |
| End\_Time | Hora de finalización del accidente |
| Start\_Lat | Latitud en coordenadas GPS del punto inicial |
| Start\_Lng | Longitud en coordenadas GPS del punto inicial |
| End\_Lat | Latitud en coordenadas GPS del punto final |
| End\_Lng | Longitud en coordenadas GPS del punto final |
| Distance | Extensión del recorrido afectado por el accidente |
| Description | Descripción del accidente |
| Number | Número de calle |
| Street | Nombre de calle |
| Side | Lado de la calle (left/right) |
| City | Ciudad |
| Country | País |
| State | Estado |
| ZipCode | Código ZIP |
| TimeZone | Timezone basado en la ubicación del accidente |
| Airport\_Code | Estación meteorológica basada en el aeropuerto más cercano al accidente. |
| Weather\_Timestamp | Timestamp del registro de observación del tiempo |
| Temperature | Temperatura en Fahrenheit |
| Wind\_Chill | Temperatura del viento en Fahrenheit |
| Humidity | Humedad en % |
| Pressure | Presión del aire (pulgadas) |
| Visibility | Visibilidad en millas |
| Wind\_Direction | Dirección del viento |
| Wind\_Speed | Velocidad del viento |
| Precipitation | Precipitación (pulgadas) |
| Weather\_Condition | Condición del tiempo |
| Amenity | Presencia de amenidad en una ubicación cercana |
| Bump | Presencia de reductor de velocidad en una ubicación cercana |
| Crossing | Presencia de cruce en una ubicación cercana |
| Give\_Way | Presencia de CEDA en una ubicación cercana |
| Junction | Presencia de intersección en una ubicación cercana |
| No\_Exit | Presencia de calle sin salida en una ubicación cercana |
| Railway | Presencia de ferrocarril en una ubicación cercana |
| Roundabout | Presencia de rotonda en una ubicación cercana |
| Station | Presencia de estación en una ubicación cercana |
| Stop | Presencia de alto en una ubicación cercana |
| Traffic\_Calming | Presencia de tráfico calmado en una ubicación cercana |
| Traffic\_Signal | Presencia de señan de tránsito en una ubicación cercana |
| Turning\_Loop | Presencia de giro en una ubicación cercana |
| Sunride\_Sunset | Periodo del día basado en sunride/sunset |
| Civil\_Twilight | Periodo del día basado en civil twilight |
| Nautical\_Twilight | Periodo del día basado en nautical twilight |
| Astronomical\_Twilight | Periodo del día basado en astronomical twilight |

La información anterior corresponde a información histórica de los accidentes reportados para el estado de **NY**.

El set de datos posee información muy detallada como la hora exacta del accidente, así como la hora en que concluyó la atención de este.

De igual forma, podemos ver atributos referentes a las condiciones climatológicas, señalización víal, entre otros

Tenemos un atributo “State” que solo contiene un valor “NY” en este caso.

Igualmente, el atributo Turning\_Loop posee un único valor “false”.

En este caso, el atributo a predecir es **Severity,** el contiene cuatro categorías diferentes que nos indican el tiempo aproximado que tardará habilitar la vía:

1: 1hrs o menos

2: 2hrs o menos

3: 3hrs o menos

4: más de 4hrs

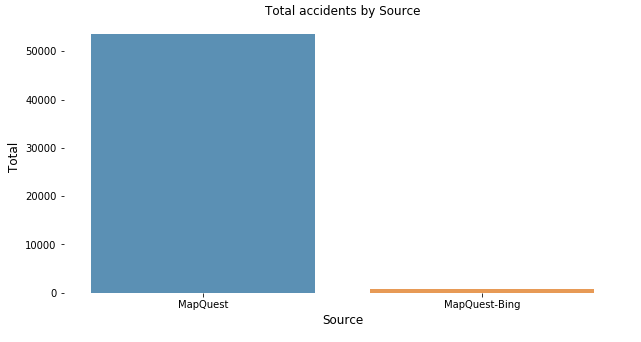
**How much confidence should stakeholders place in the results?**

El modelo será construido basado en los datos que representan el historial de accidentes reportados en un periodo en particular para el estado en investigación (NY), por lo que la severidad del accidente reportada por el modelo será de suma confianza y podrá ser usada en el proceso de gestión de recursos, al momento de atender un accidente.

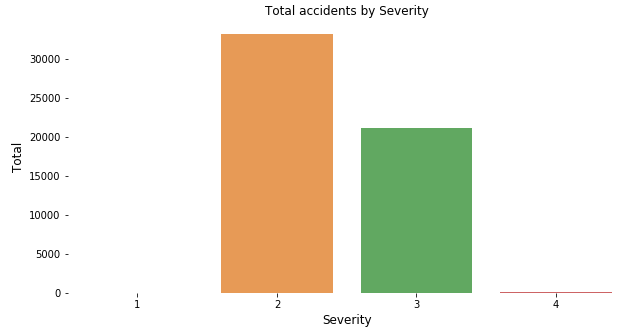
# Data Visualization

A continuación, algunas de las visualizaciones logradas con los datos:

Accidents by Source

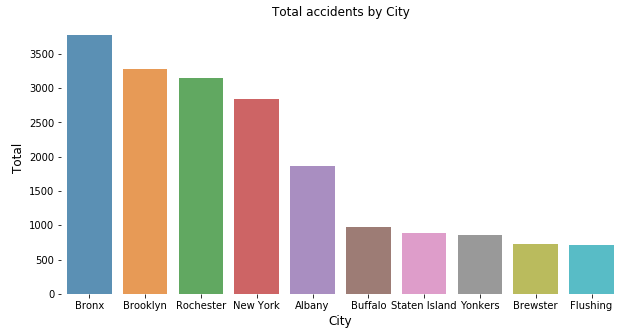


Accidents by Severity

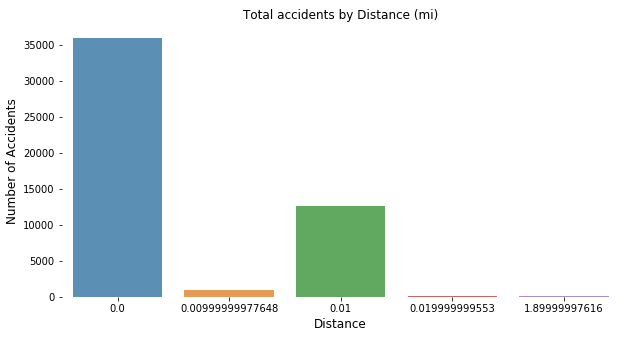


Como Podemos observar en la imagen anterior, las severidades tipo **2 y 3** son las que se presentan con mayor frecuencia.

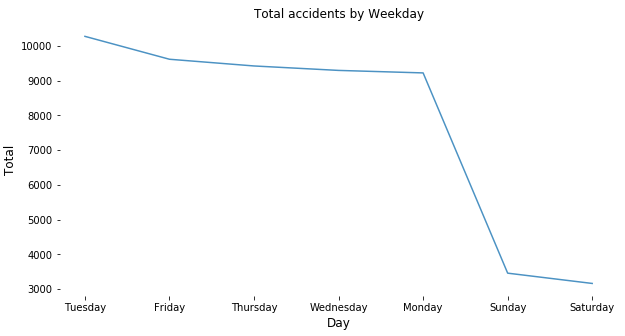
Accidents by City



Accidents by Distance

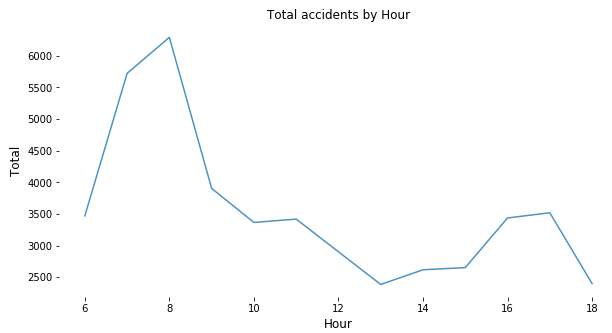


Creamos una nueva columna para mostrar los accidentes agrupados por día de la semana:



Podemos ver que la mayor cantidad de accidentes son reportados durante los **días laborales**, mientras que el fin de semana reporta una disminución en el reporte de accidentes.

Ahora, veamos los accidentes agrupados por hora:



Las horas con mayor reporte son de **6am - 9am,** que coincide con la hora en que la mayoría de las personas se desplaza a sus trabajos.

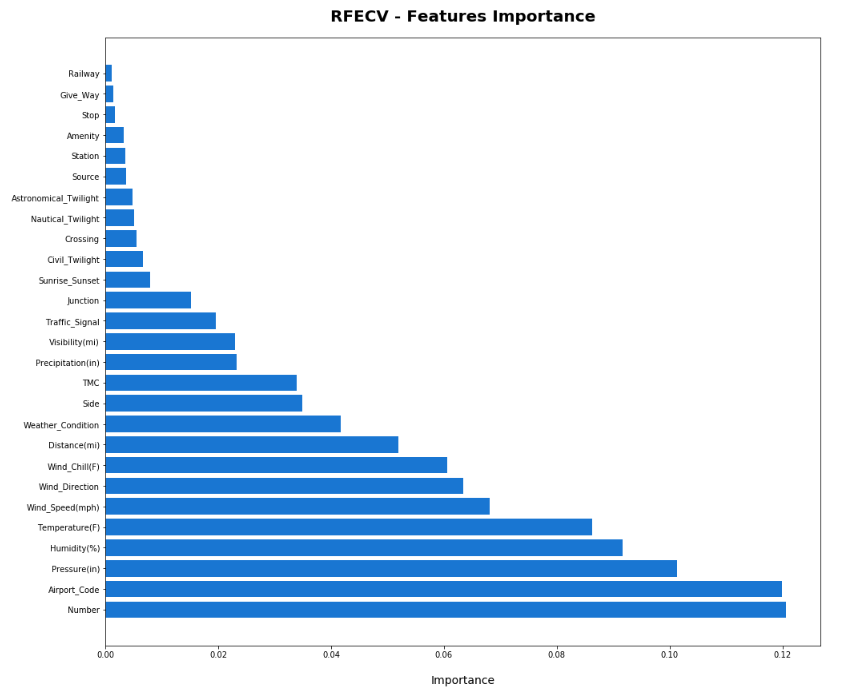
# Pre-processing and feature engineering

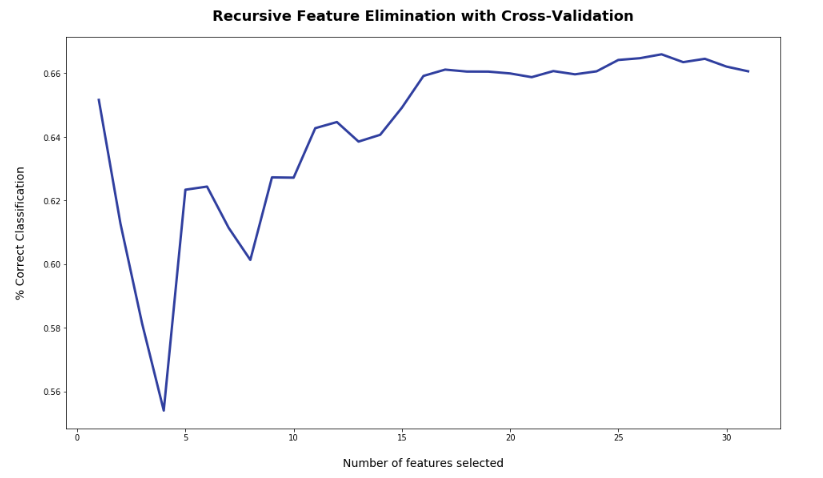
La información para crear el modelo ha sido facilitada por los stakeholders. Dicha información corresponde a información histórica de los accidentes reportados para el estado de NY.

Se hizo una revisión a cada atributo con el fin de asegurarnos que posee el tipo de dato correcto. También se evaluó cada atributo con el fin de determinar si aportan valor o no al modelo y eliminamos aquellas que consideremos innecesarias.

**Which attributes in the data can we deem to be statistically significant to the problem at hand?**

Basado en el análisis realizado sobre los datos con los que disponemos, los siguientes son los atributos más relevantes para definir la severidad de un accidente. Los mismos están ordenados por importancia:





De acuerdo con las imágenes anteriores, para el entreno de los modelos se usaron las siguientes variables debido a que son las que representan mayor importancia:

* Distance(mi)
* Number
* Airport\_Code
* Temperature(F)
* Wind\_Chill(F)
* Humidity (%)
* Pressure(in)
* Wind\_Direction
* Wind\_Speed(mph)
* Weather\_Condition
* Bump
* No\_Exit

También aplicamos la técnica de correlación, la cual mostró que los siguientes atributos están altamente relacionados: Traffic\_Calming, Wind\_Chill(F). Se decidió eliminar el atributo **Traffic\_Calming** del set de datos.

De igual forma se excluyeron los tributos **Start\_Time, End\_Time,** que analizando el set de datos descubrimos que tienen una correlación directa con la variable dependiente.

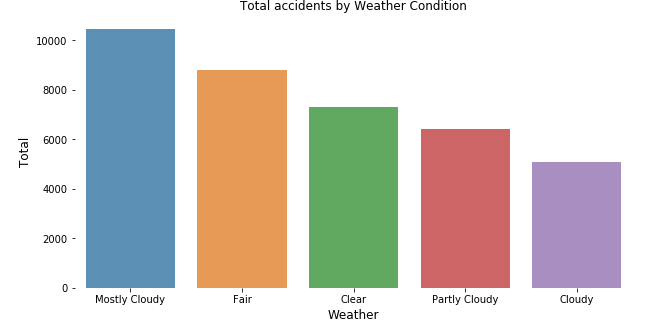
Los atributos **Roundabout** y **State** fueron descartados ya que presentaban un único valor (false y NY respectivamente) para todos los registros.

# Hyphotesis

La hipótesis definida inicialmente es la siguiente:

*A peores condiciones climatológicas, mayor número de accidentes reportados.*

Después del análisis, podemos concluir que las **condiciones climatológicas** no favorables, no necesariamente indican un incremento en la cantidad de accidentes, tal y como lo habíamos planteado. Podemos ver que inclusive con buenas condiciones climatológicas también tenemos una cantidad similar de accidentes reportados. Lo podemos ver en el siguiente gráfico:



# Tool and Libraries

El proyecto fue desarrollado haciendo uso de la herramienta Anaconda, sobre el lenguaje de programación **Python**.

Algunas de las librerías utilizadas para modelar los datos, así como la exploración y preprocesamiento son las siguientes:

* Pandas
* Numpy
* Matplotlib
* Seaborn
* sklearn

# Models

Los siguientes modelos fueron entrenados usando diferentes parámetros con el objetivo de encontrar el que mejor exactitud nos ofreciera para el problema en estudio:

1. RandomForestClassifier
2. Support Vector Classifier
3. KNeighborsClassifier

# Accuracy and Cross Validation results

A continuación, las métricas para cada uno de los modelos después de entrenarlos:

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | CV Score |
| RandomForestClassifier | **0.78** | **0.83** |
| KNeighborsClassifier | 0.61 | 0.67 |
| Support Vector Classifier | 0.68 | 0.79 |

Basado en las métricas anteriores, y en el performance mostrado al momento de la ejecución, el modelo seleccionado para realizar la predicción es el **RandomForesrClassifier.**

**Data Splitting**

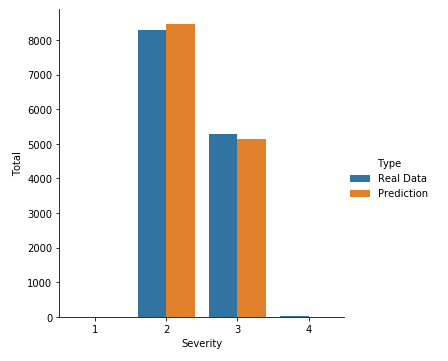
El 75% de los datos se usó para entrenar el modelo y el 25% restante fue usado para probar el modelo:

|  |  |
| --- | --- |
| Type | Records |
| Training | 40838 |
| Testing | 13613 |

# Predictions

Las siguientes son las predicciones obtenidas con el modelo seleccionado:

|  |  |  |
| --- | --- | --- |
| Severity | Type | Total |
| 1 | Real Data | 3 |
| 2 | Prediction | 8461 |
| 2 | Real Data | 8292 |
| 3 | Prediction | 5152 |
| 3 | Real Data | 5296 |
| 4 | Real Data | 22 |



Como podemos ver, de un total de 13613 registros de accidentes, **13475** registros de la predicción acertaron los datos reales. Representado en forma de porcentaje, es un **97.51**% de exactitud.

# Lessons Learned

* Con respecto al set de datos y la hipótesis que planteamos al inicio del proyecto, nos encontramos con particularidades inesperadas como, por ejemplo, que la información referente a la condición climatológica no representa un factor tan determinante con respecto al número de accidentes reportados, para este caso en particular**.**
* El facilitar a los conductores información referente al tiempo que demorará una vía cerrada, permite que estos puedan tomar mejores decisiones en lo que respecta a la ruta a tomar, evitando repercusiones en el ámbito laboral y familiar.
* A pesar de que el porcentaje de exactitud mostrado por el modelo no es el más favorable, el porcentaje final de registros de la predicción que coinciden con la información real fue muy satisfactorio.

# Recommendation

* Una vez concluido el modelo y con los resultados obtenidos, mostraremos diferentes visualizaciones a los interesados, donde se pueda comparar el porcentaje de accidentes clasificados con severidad **alta, media y baja**.
* Tomar en cuenta los diferentes factores que influyen en esta clasificación, con el fin de crear un mayor conocimiento que permita la agilización de recursos, tanto humanos como médicos al momento de atender un accidente.
* Usar la información generada por este modelo para bien de los usuarios conductores que día a día se enfrentan con dificultades viales en el camino a sus trabajos. Esta información gestionada de manera correcta puede resultar muy útil.