# Analytics Startup Plan

**Synopsis: *This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

|  |  |
| --- | --- |
| **Project** | Toronto Traffic Collision Prediction Model |
| **Requestor** | Centennial College |
| **Date of Request** | July 11, 2023 |
| **Target Quarter for Delivery** | August 16, 2023 |
| **Epic Link(s)** | N/A |
| **Business Impact** | The business impact of this project is to help build safer streets for the residents and road users of Toronto by minimizing traffic collisions in the city. The project will support the Toronto Traffic Services in identifying hot spots on Toronto roads where there is high risk of traffic collisions occurring, as well as provide a deeper understanding of the key factors that lead to an increased risk of collisions.  The project will also allow traffic services to take proactive actions based on the identified risky road segments/intersections to mitigate collision risk, inform road design and public policy choices, as well as assist in the optimization of police resource deployment. |

## 1.0 Business Opportunity Brief

|  |  |
| --- | --- |
|  | Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment. |

In the last 10 years, a total number of 1,581 deaths have resulted from road accidents in Toronto,

Canada between 2011 and 2020 with an average yearly road accident-related death of 158.1. Road accidents are the 3rd leading cause of death in Toronto.

The project arose from the need to provide data-driven insights and recommendations in delivering safety measures to reduce road accidents for the Vision Zero 2.0 project which aims to eliminate collisions in several world cities, including Toronto. An important part of the Vision Zero 2.0 project is to minimize collisions through policy and road design changes, and strategic deployment of police resources to high-risk roads and intersections.

This project will assist in the identification of roads and intersections with high collision risk by developing a traffic collision prediction model for the City of Toronto based on various data points – past collisions, weather data, average traffic speeds and volumes, etc, to predict the likelihood of a collision occurring on a particular road segment on a given day. Using these likelihood estimates, roads and intersections can be ranked or mapped according to collision risk, which will allow traffic services to determine where best to dedicate scarce police resources, as well as identify roads where design or policy changes are required to decrease the risk of collisions before they occur.

## 1.1 Supporting Insights

|  |  |
| --- | --- |
|  | Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally? |

Similar collision prediction models developed in the past:

1. City of Montreal: [1905.08770.pdf (arxiv.org)](https://arxiv.org/pdf/1905.08770.pdf)
2. State of Utah: [Using Machine Learning to Predict Car Accident Risk | by Daniel Wilson | GeoAI | Medium](https://medium.com/geoai/using-machine-learning-to-predict-car-accident-risk-4d92c91a7d57)
3. New York City: [Predictive Analytics on NYC Collision Data | by Anushka Sandesara | Medium](https://anushkasandesara.medium.com/predictive-analytics-on-nyc-collision-data-9f06c94140f2)

## 1.2 Project Gains

|  |  |
| --- | --- |
|  | *Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI’s.* |

The successful completion will allow traffic services to determine where best to dedicate scarce police resources, as well as identify roads where design or policy changes are required to ensure safety for users.

If sufficient precautions are taken by the city to mitigate collision risks identified by the model, it should lead to a decrease in the number of collisions on Toronto roads leading to reduced loss of life and property, as well as increased operational efficiency for the Toronto police services.

The implications of not carrying out this project would lead to preserving the status quo of high number of traffic collisions in the City of Toronto, leading to loss of life and property, as well as failure to progress on the objectives of the city’s Vision Zero plan of reducing traffic collisions and fatalities.

## 2.0 Analytics Objective

|  |  |
| --- | --- |
|  | List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.  Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed. |

The primary objective of this project is to develop a collision prediction model to estimate the probability of a traffic collision occurring on a given road segment or intersection within the City of Toronto in a given day.

## 2.1 Other related questions and Assumptions:

|  |  |
| --- | --- |
|  | *List any assumptions that may affect the analysis.* |

We will also look at these specific to gain a better understanding of traffic collisions in Toronto:

1. Which intersections and roadways in Toronto have the highest collision rates?
2. When and where to most crashes occur?
3. How does the spatial pattern of fatal collisions differ from the spatial pattern of collisions overall?
4. Where are collisions likely to disproportionately affect vulnerably road users?

The major assumption made by this project is that the likelihood of traffic collisions occurring can be modled and predicted by analyzing the features of a road segment or intersection such as traffic and pedestrian volume, traffic controls, and previous collision record.

## 2.2 Success measures/metrics

|  |  |
| --- | --- |
|  | *What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.* |
|  | The success metrics used to evaluate the performance of the predictive models will be – accuracy, precision, and recall.  For the project to be considered successful, it must be able to identify road segments/intersections within the City of Toronto that have high risk of traffic collisions. These road segments and intersections must be rank-able by the predicted collision probability, or the predicted number of collisions, so that the Toronto traffic services can identify which intersections to dedicate resources and design changes towards. |
|  |  |

## 2.3 Methodology and Approach

|  |  |
| --- | --- |
|  | *Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.* |

**Type of Analysis:** Decision Tree, Neural Networks, Random Forest, Gradient Boosting (XGBoost).

**Methodology:** The methodology of this project will consist of the following project steps:

1. Gather datasets.
2. Merge all datasets together.
3. Conduct exploratory analysis for modeling.
4. Data pre-processing.
5. Conduct descriptive analytics.
6. Feature Engineering.
7. Modeling.
8. Performance review and hyper-parameter tuning.
9. Create risk map of roads/intersections.
10. Interpretation and conclusion.

We will start by gathering all the relevant datasets for the project – Toronto All Collisions, KSI Data, Traffic and Pedestrian Volumes, Average speeds at roads, Weather data etc. and merge them together into a single dataset.

Then, we will conduct exploratory analysis of the data to identify and treat missing values, outliers, standardize numerical values etc. Then, we will conduct descriptive analytics and data visualization to gain a deeper insight into the data structure and the problem using Python or Tableau. At this stage, we might also conduct unsupervised analysis such as k-means clustering to understand the data structure further.

Next, we will prepare the data for modeling and carry out feature engineering (if needed) from the variables identified as important from the descriptive analysis. The final training dataset will all be assigned a target of 1 – meaning a collision occurred. These are the positive samples that we will be using for our modeling.

Since we only have data for instances where a collision occurred, we will have to create the negative samples – instances where a collision did not occur. Any time and road segment combination where a collision did not occur can be considered a negative sample. The negative samples will be derived from randomly sampling from the positive samples and randomly altering the time of day, the day of month, the road segment, or any other variables. These will be assigned a target of 0. Finally, the positive and negative samples will be combined into the final training dataset.

We will build multiple machine learning models to predict collision risk - decision trees, random forests, neural networks, gradient boosting etc. We will also conduct PCA on the dataset and build models from the principal components. Finally, the performance of these models will be compared using accuracy, precision, and recall statistics to select the best model.

**Output:** The output will be an interactive map of Toronto roads and intersections, with identified hot spots of high-risk zones for traffic collision, color coded by predicted risk. The map will be derived from the predictions of the best model identified by the performance metrics.

**Actual Project Methodology**

# **Exploratory Data Analysis (EDA)**

## **Initial Data Exploration**

Once the merged dataset is prepared and exported, we conduct EDA on the dataset to prepare for modeling. We read the merged dataframe as a GeoDataFrame called final\_df.

### **Missing Values**

We checked the merged dataset for missing values. If found, missing values can be imputed. However, the dataset does not contain any missing values.

A screenshot of a computer program

Description automatically generated

### **Duplicate Rows**

We also check the dataset for duplicate rows based on the centreline\_id that may have been created during the merging process. We find that the data contains 5,993 duplicate rows. The duplicate rows are dropped, and the resulting data frame has 7,986 rows.

A screen shot of a computer program

Description automatically generated

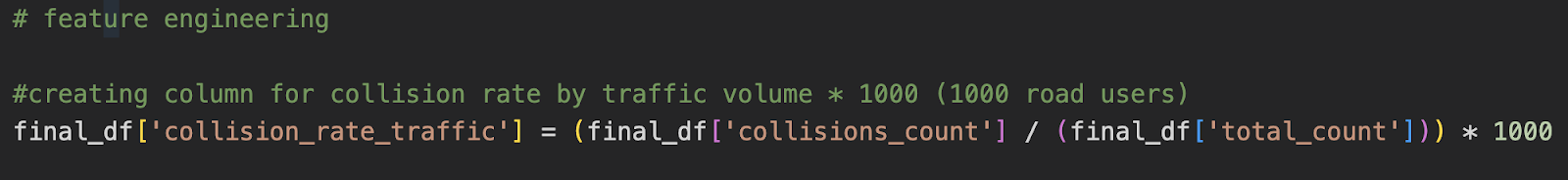
## **Feature Engineering**

Feature Engineering is the process of creating new and meaningful features from existing data to improve the performance of a machine learning model. For our analysis, we engineer three new features: Collision Rate, Collision Class, and Speed Hump Present.

### **Collision Rate**

The Collision Rate feature was derived due to the fact that road segments that have higher traffic volumes are likely to have higher collision numbers, even if the design of the road segment is not particularly dangerous. As we want to understand how the design and infrastructure of a road segment can affect the collision risk, and the differences between road segments that have high versus low collision risk, we will be using the *collision rate* of that road segment for our model.

The collision rate is derived by dividing the collisions\_count of a road segment by the total daily volume of traffic. The collision rate is then multiplied by 1000 to get the collision rate per 1000 road users (including vehicles, trucks, pedestrians, and cyclists).



### **Collision Class**

The collision class feature was derived from collision counts to bin the total number of collisions for use in classification models. The total collisions count was divided into 5 bins: less than 5 collisions, 5 - 20 collisions, 20 - 50 collisions, 50 - 80 collisions, and over 80 total collisions count.

A screen shot of a computer

Description automatically generated

### **Speed Hump Present**

This feature was derived to use the speed hump count as a boolean flag indicating whether a speed hump was present or not in a given road segment. Speed humps were chosen as they are by far the most common traffic calming measure found in our dataset.

A black screen with white text

Description automatically generated

## **Correlations**

After the feature engineering step, we wanted to look at the correlation heatmap of all the existing and engineered features in the dataset. As shown in the heatmap below, outside of the traffic volumes and percentage columns, no features in the dataset have significant correlation with each other.

A screenshot of a computer screen

Description automatically generated

## **Data Processing**

### **Dropping Columns**

The traffic volumes and percentage columns in the dataset were correlated with each other. Therefore, we drop the percentage columns, keeping only the traffic volumes columns. We also dropped the total count column, as the information in the column was captured in the count column for other transportation modes. Additionally, we drop some other columns that are not required for the analysis or highly imbalanced- such as road name, speed cushion count, road hierarchy etc.

A computer screen with text

Description automatically generated

### **Setting Data Types**

We set the data types for the variables to the proper format for modeling, including setting the data type of categorical variables to “category” and collisions count and number of lanes from “float” to “int64”.

A screen shot of a computer screen

Description automatically generated

### **Skews and Outliers**

The numerical variables in our dataset were highly skewed and contained significant outliers, as shown in the figure below. Skewed data distributions can negatively impact the performance of machine learning models and hinder their ability to learn from the data. They can also bias the models understanding of how features impact the target variable, with models assigning higher importance to features with a larger range of values. Thus, we needed to treat the skewed distributions with cap & floor using a range of 3IQR, as well as log transformation techniques.

A screenshot of a computer

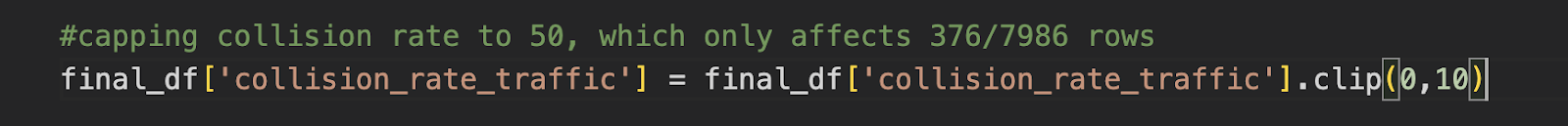
Description automatically generated

#### **Collision Rate**

Collision Rate was the variable with the highest skew. We capped the collision rate at 10, as the data distribution showed that although the maximum collision rate was 383, 75% of the road segments have collision rates of under 3. Even with capping the collision rate using a 3 IQR at 10, the data still contained some outliers, as shown in the box plot. We decided to leave those outliers to preserve the information about road segments with high collision rates. Later, we create a separate dataframe of the road segments with these outlier collision rates to model them separately.

A screenshot of a computer

Description automatically generated



A graph of a graph showing a graph

Description automatically generated

#### **Collisions Count**

Collisions Count was another target variable that we wanted to use that had a relatively high skew of 4.322 and thus had to be treated.

A screenshot of a computer program

Description automatically generated

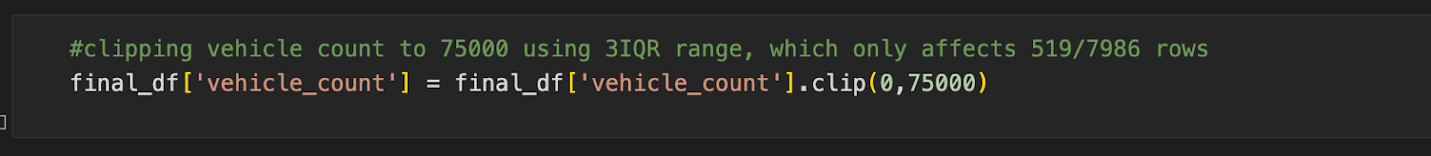
The descriptive statistics of the variable shows that the variable had significant outliers, with a minimum collision count of 0 and 75th percent of the data being under 40. However, the maximum count was 1064. Using the pandas .clip() function, the collision count was capped using a range slightly higher than 3 IQR range at 150 collisions. In our modeling step, we will be using a collisions count of over 80 as a very high collisions count.

The resulting box plot shows that the data still has quite a few outliers, but we decided to keep them so as not to lose valuable information in our dataset.

#### **Vehicle Count**

A screenshot of a computer

Description automatically generated



As shown by the descriptive statistics and histogram, vehicle count was also relatively skewed, with a long tail on the right. Again, we capped this variable using a range of 3 IQR, making sure not to affect too many rows.

#### **Trucks Count**

A screenshot of a computer

Description automatically generated

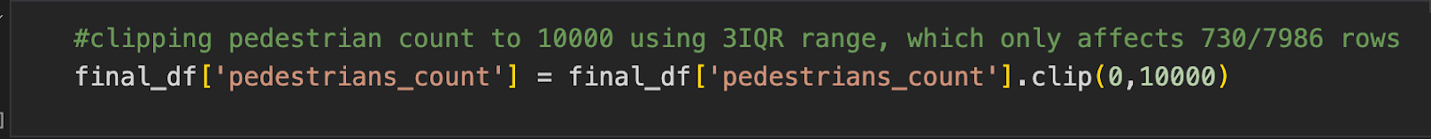
A screen shot of a computer

Description automatically generated

Trucks Count was capped at 3000 using a 3 IQR range.

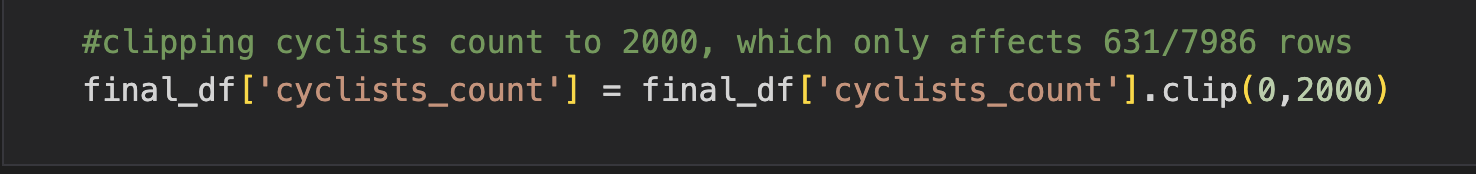
#### **Pedestrian Count**

Pedestrian Count was capped at 10,000 using a 3 IQR range.



#### **Cyclists Count**

Cyclists Count was capped at 2000



### **Log Transformation**

After conducting Cap & Floor on our numerical variables to treat the outliers, we performed Log Transformation on our dataset to fix the skewed distributions, as well as bring the dataset values to a similar range, using a log base of 2. Log Transformation is used to reduce skewness in the dataset by compressing the range of values, making the distributions more symmetric and bringing extreme values closer to the mean. Log Transformation was chosen instead of other methods as our data is mostly positively skewed and contains outliers.

As some of our numerical variables contained values that were 0, we could not apply Log Transformation directly. Therefore, we added 1 to all numerical values in the dataset before applying the log transformation.

A screen shot of a computer program

Description automatically generated

Upon re-evaluating the skews in the dataset, we found that the log transformation had successfully fixed the skewed distributions.

A screen shot of a computer code

Description automatically generated

The box plots of the numerical variables also show that the data is symmetrical, and all variables are in a similar range, with only vehicle count having noticeable outliers.

A diagram of different colored squares

Description automatically generated

### **Summary**

In summary, we started our EDA on the dataset by checking for missing values and duplicate rows, which were dropped. After that, we conducted feature engineering, creating three new features: collision rate, collision class, and speed hump present. The collision rate and collision class features will be used as targets for our regression and classification models respectively. We then checked for correlations between variables to decide which variables to include in our modeling. We discovered our dataset did not contain many correlated variables, except for the traffic counts and percentages.

Therefore in the data processing step conducted next, we dropped the traffic percentage columns, along with other columns that were not necessary to our analysis, and also set the data types of the remaining variables to the correct format.

Finally, we treated the skewed variables and outliers in our dataset by applying cap & floor, as well as log transformation to the skewed variables, and bringing them to a similar range. The final skews and box plot showed that the treated variables all had skews close to zero and were on a similar range. Therefore, we decided not to perform additional treatment on the data set such as standardization or normalization.

## 3.0 Population, Variable Selection, considerations

|  |  |
| --- | --- |
|  | Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today. |

**Audience/population selection:**

**Observation window:** Although the data set contains data since 2008, we will only be taking observations from the past 7 years, so 2016 onwards.

**Inclusions:** All variables in the dataset will be included.

**Exclusions:** N/A

**Data Sources:** Toronto Police All Collisions, Toronto Police KSI, Toronto weather data, Traffic and Pedestrian Volumes at intersections, Average speed of roads

**Audience Level:** Toronto Police Traffic Services

**Variable Selection:**

**Derived Variables:**

**Assumptions and data limitations:**

## 4.0 Dependencies and Risks

|  |  |
| --- | --- |
|  | Identification of key factors that may influence the outcome of the project and likelihood of it happening: |

|  |  |  |  |
| --- | --- | --- | --- |
| **Risk** | **Likelihood (based on historical data)** | **Delay (based on historical data)** | **Impact** |
| *Churn rate being inflated by counting multiple contracts from the same rooftop as individual observations.* | *Low* |  | *Once analysis begins, we can quantify the inflation. However, this approach allows us to compare how the same dealer performed across different contracts and find useful patterns.* |

## 5.0 Deliverable Timelines

|  |  |
| --- | --- |
|  | List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Item** | **Major Events / Milestones** | **Description** | **Scope** | **Days** | **Date** |
| 1. | Kick-off / Formal Request |  |  |  | July 11, 2023 |
| 2. | Analysis Plan |  |  |  | July 24, 2023 |
| 3. | Gather and Merge Datasets | Merge multiple datasets into one. |  |  | July 25, 2023 |
| 4. | Data Exploration & Analysis   * Issues with duplicates * Issues with Spend data | Data Exploration + Descriptive Analysis/Data Visualization to understand data structure. |  |  | July 28, 2023 |
| 5. | Data Pre-Processing and Feature Engineering | Pre-processing data for modeling + Feature engineering new model features. |  |  | July 31, 2023 |
| 6. | Modeling | Modeling and Model Performance Review + Hyperparameter Tuning. |  |  | August 7, 2023 |
| 7. | Model Deployment | Deployment of model predictions as an interactive map. |  |  | August 10, 2023 |
| 8. | Governance and Documentation | Document model governance and steps. |  |  | August 14, 2023 |
| 9. | Stakeholder Presentation | Develop stakeholder presentation about model findings. |  |  | August 16, 2023 |
| 10. | Delivery & sign-off | Final delivery of project. |  |  | August 25, 2023 |