UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Toronto Fire Incident

Feature Engineering for Machine Learning & AI

Mabelly Tuchsznajder | Mehrdad Mlekmmohammadi

Our goal is to improve a classifier performance using feature engineering

We used the Toronto Fire Incidents data set

to predict if an incident had a civilian casualty or not

Data sets

Toronto Fire Incidents

contains information about incidents, alarm system and sprinklers on the premises, type of building/business, and impact the incident

Stations

each Toronto station with latitude and longitude information. Used to get the nearest station from the incident site

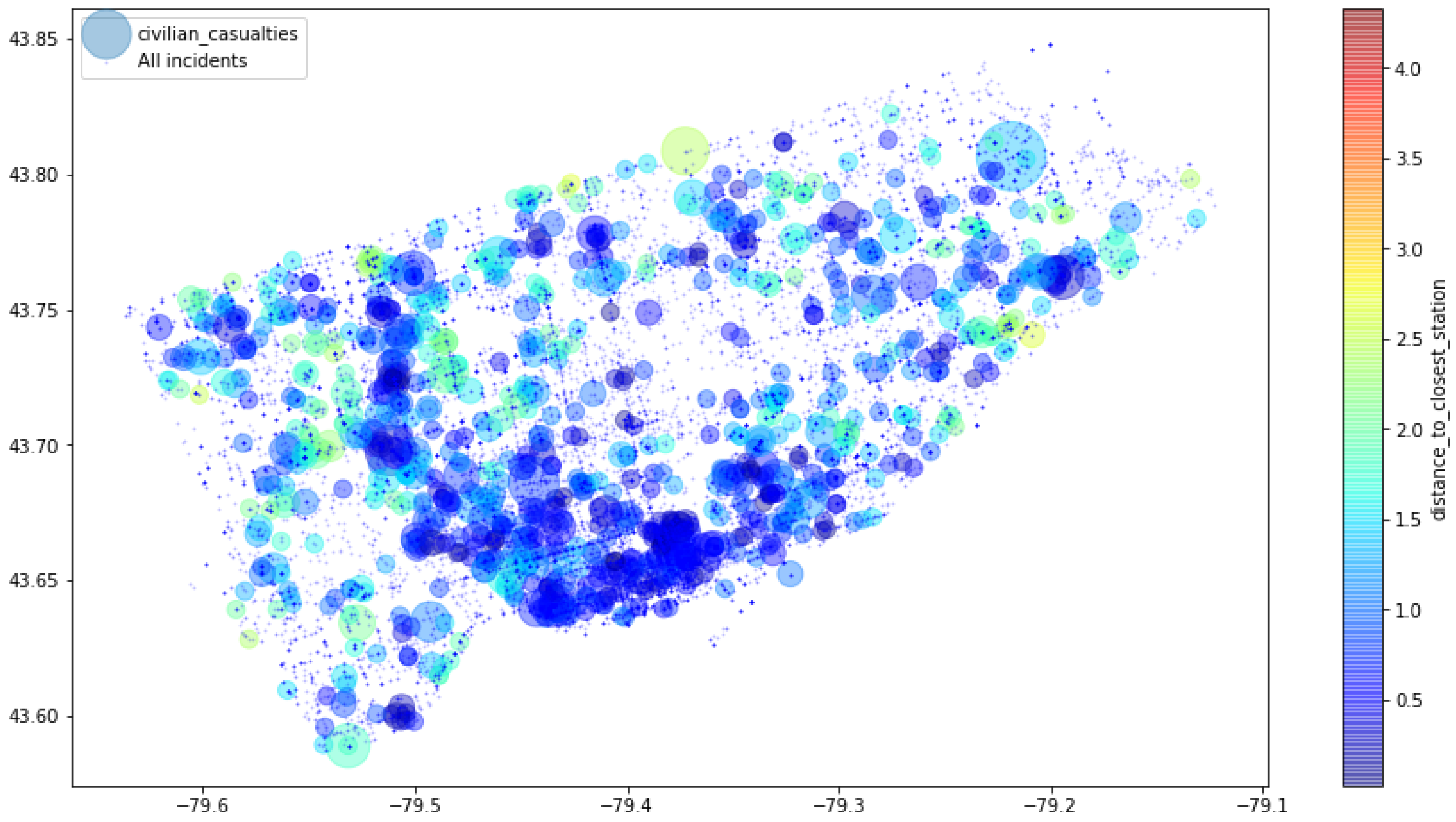
Hydrants

similar to the station data set, contains geographic information. Used to get the nearest hydrant from the incident site

Ward profile

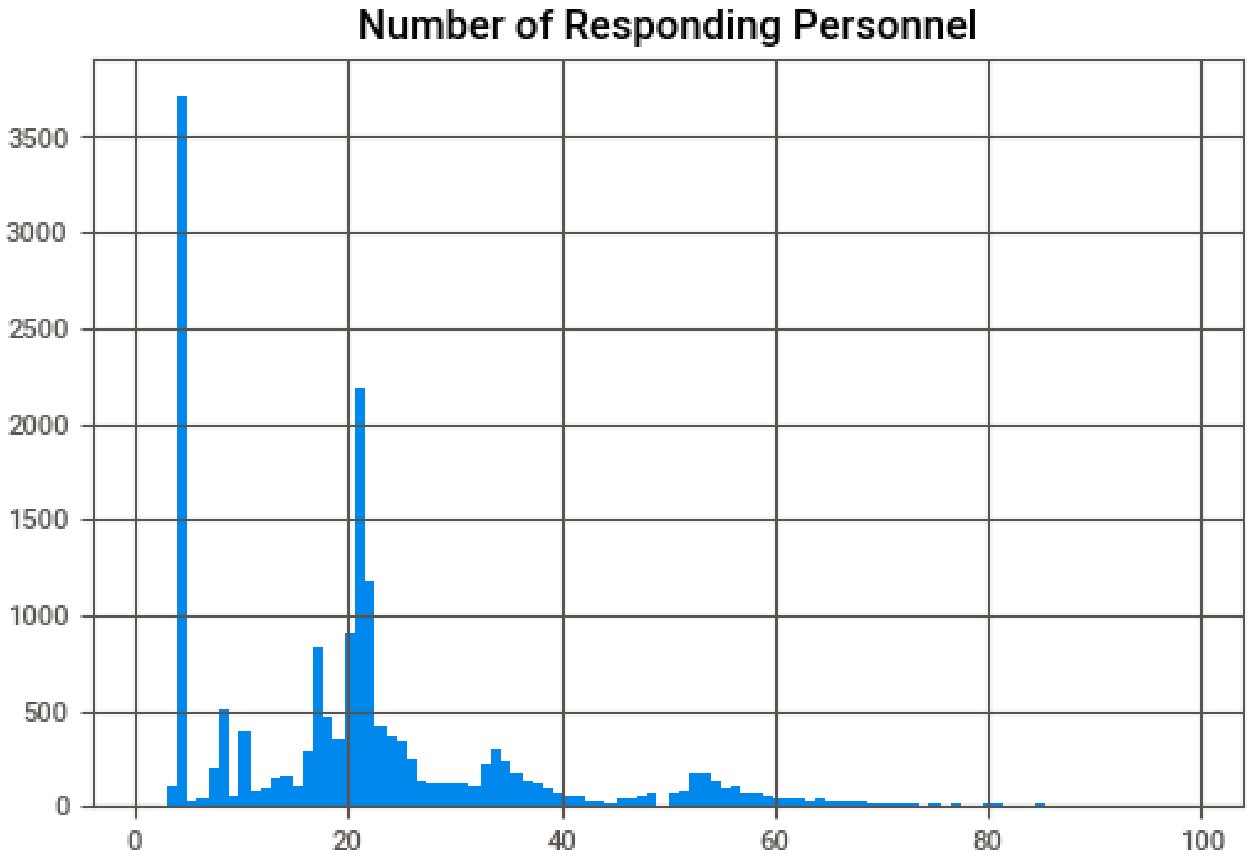
used to get information on population density and number of building by type

EDA: Civilian casualties



About 6% of the incidents result in civilian casualties

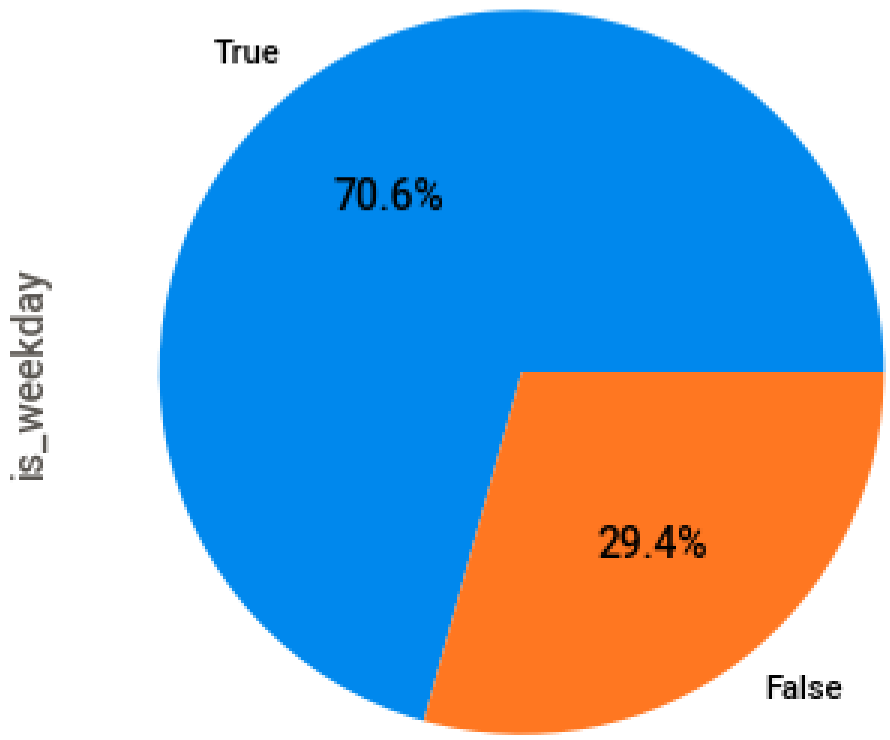
EDA: Number of responding personnel



It seems that the number of responding personnel has two peaks at 4 and 21

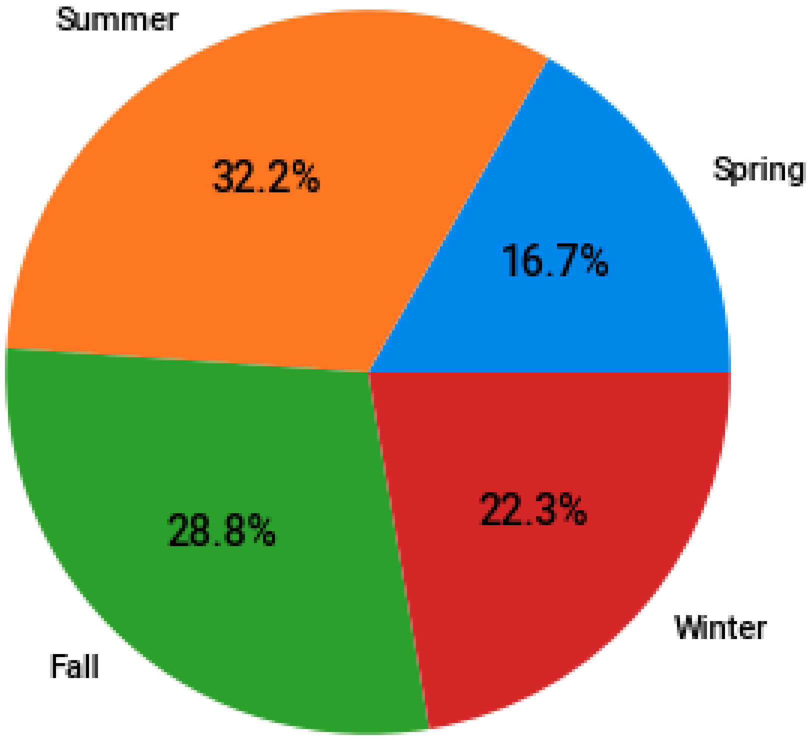
The incidents with less than 5 personnel involved (23% of the incidents) contribute to a very small portion of casualties (less than 2%)

EDA: Time of incidents (day and season)



Day of week

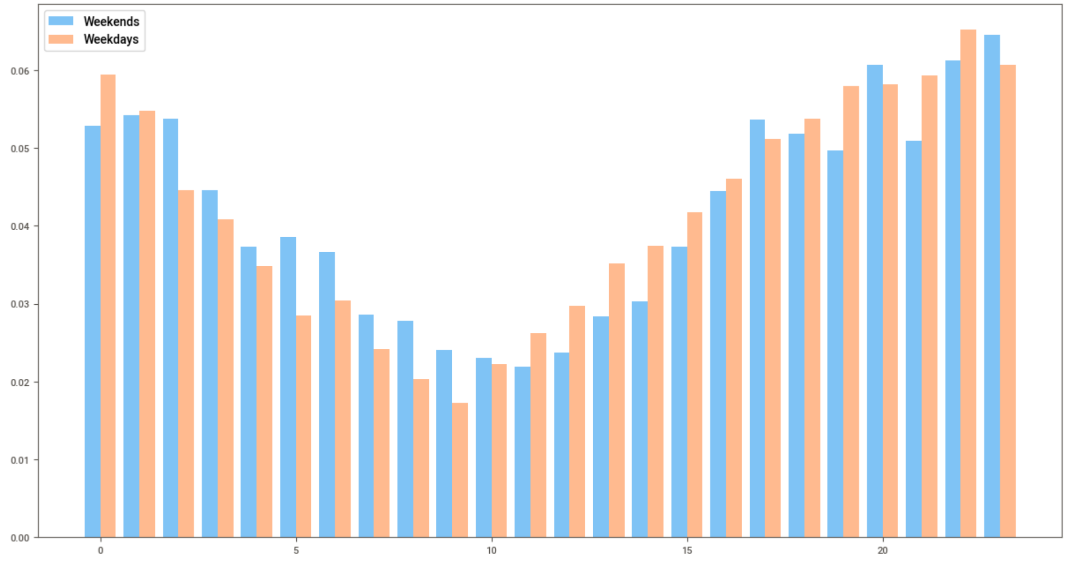
Weekends account for a slightly larger number of incidents



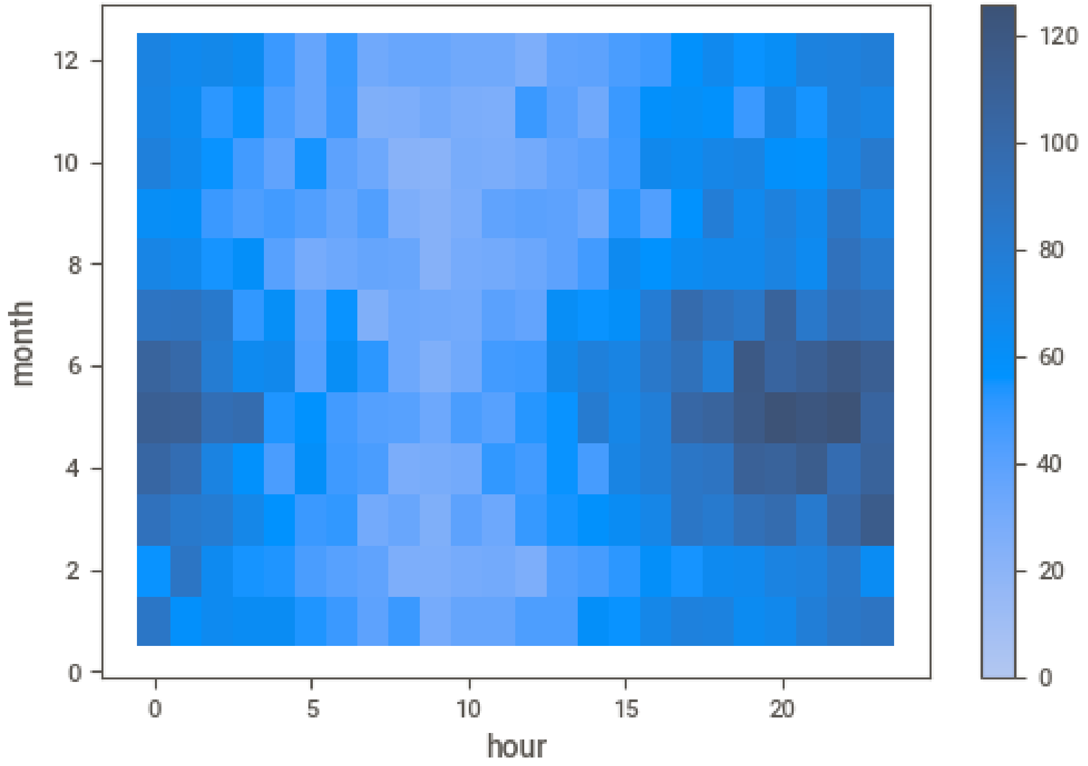
Season

most incidents happen in summer and Fall

EDA: Time of incidents (hours)

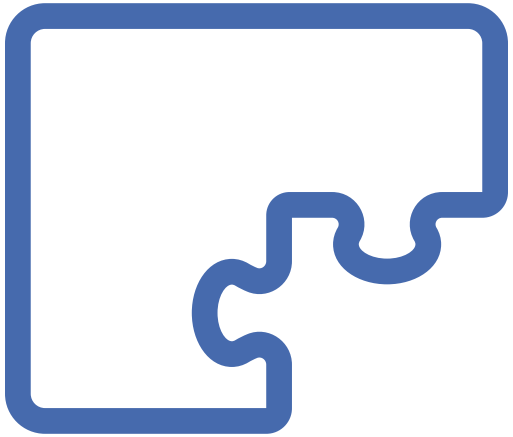


Hour of incidents



Hour of incidents within the months

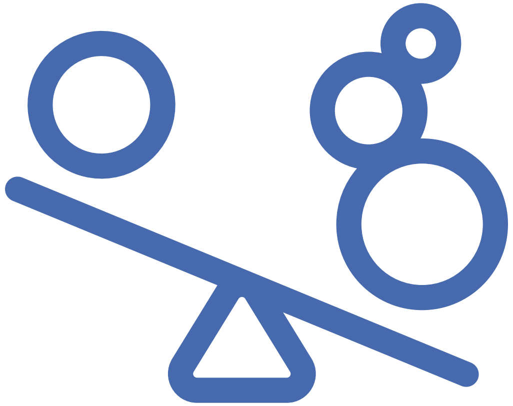
Several issues were identified during the EDA



Missing Values

- **~70%** of the features had missing values
- **How was it solved?** deleting features with more than 50% of missing values, filling NA's on categorical and numerical values based on *initial event type*

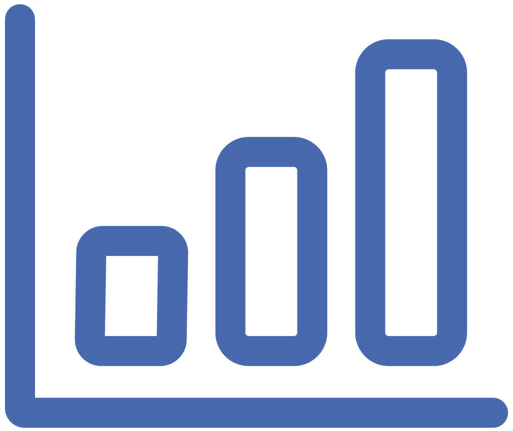
Several issues were identified during the EDA



Imbalanced data

- **What?** only **6%** of the incidents had causalities
- **How was it solved?** using **class-weight** parameter

Several issues were identified during the EDA



Categorical data

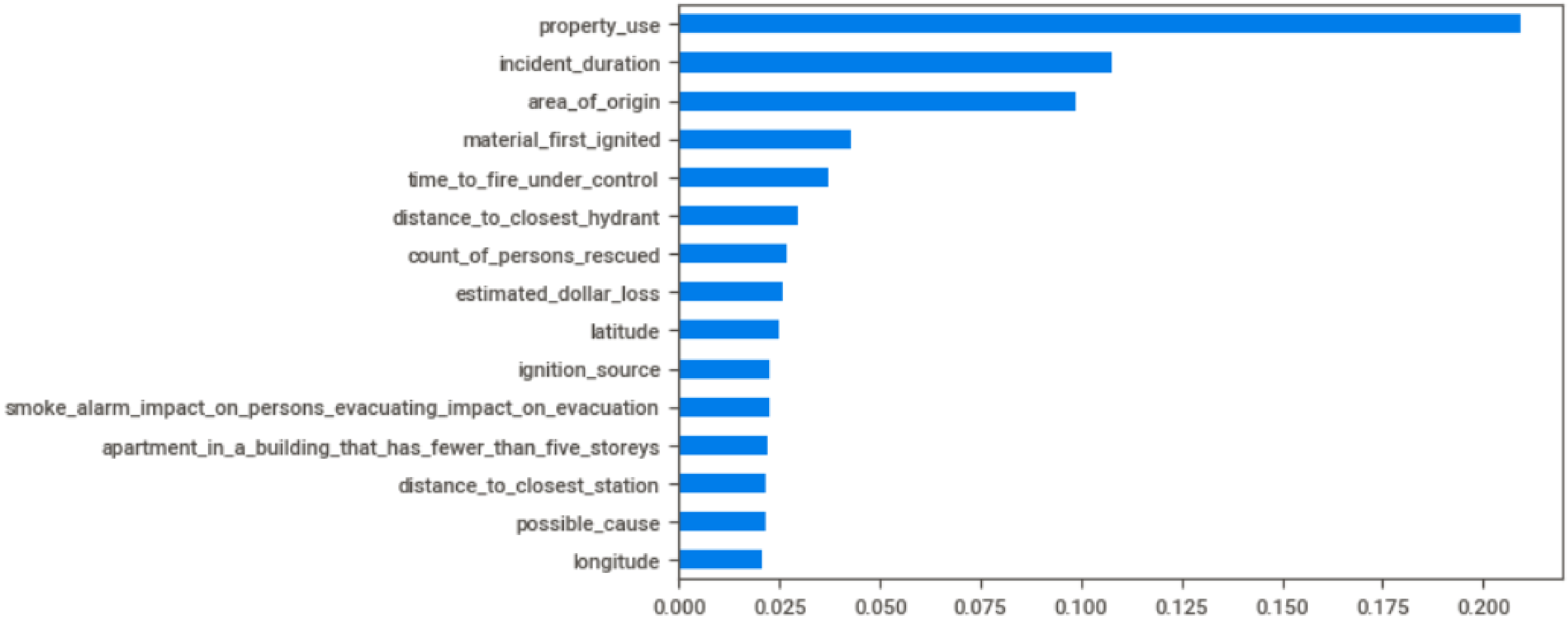
- **What?** most of the features on the Toronto Incidents data set were categorical
- **How was it solved?** applying Weight of Evidence (WOE)

New features added were added to improve the model score

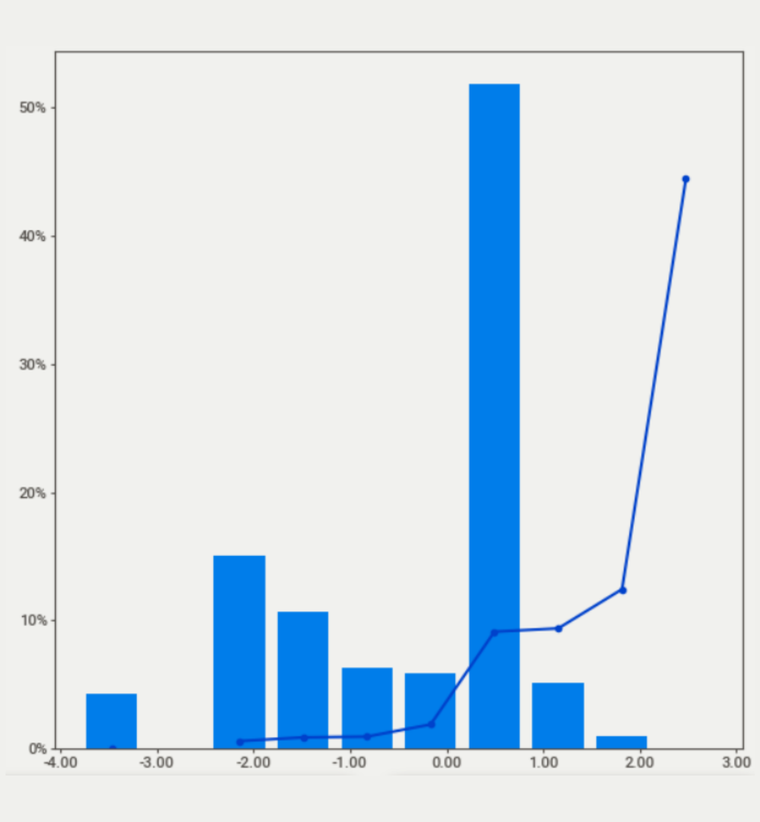
- Response time
- Incident day of the week
- Population density on the ward
- Minutes until the fire was under control
- Month when the incident happened
- Distance to the nearest Fire Station
- Duration of the incident
- Year of the incident
- Distance to the nearest hydrant

Feature Importance Analysis

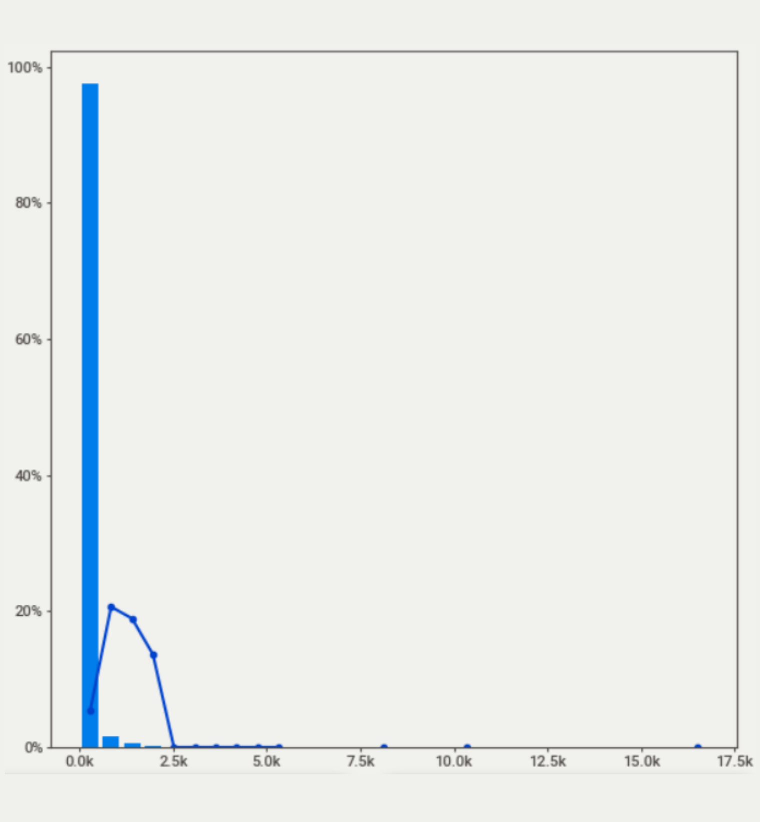
~30% of the top 15 features are features that came from **another data set** or was **created based on other features**.
About 45% of the new features created are among the top 15 most important features.



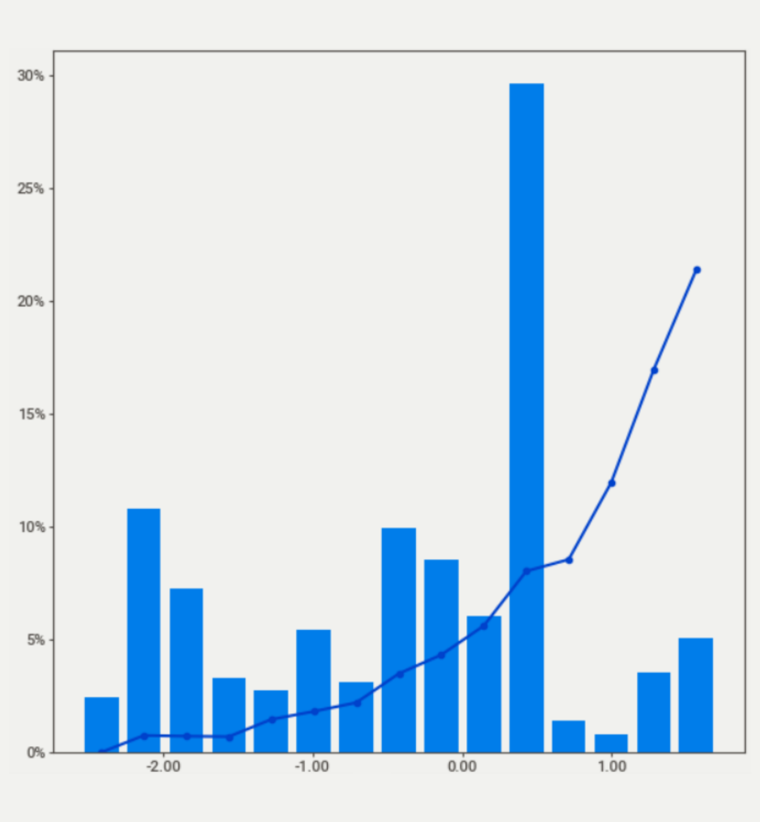
The top 3 most important features were responsible for 42% of the model variance



Property Use (WoE)



Incident Duration

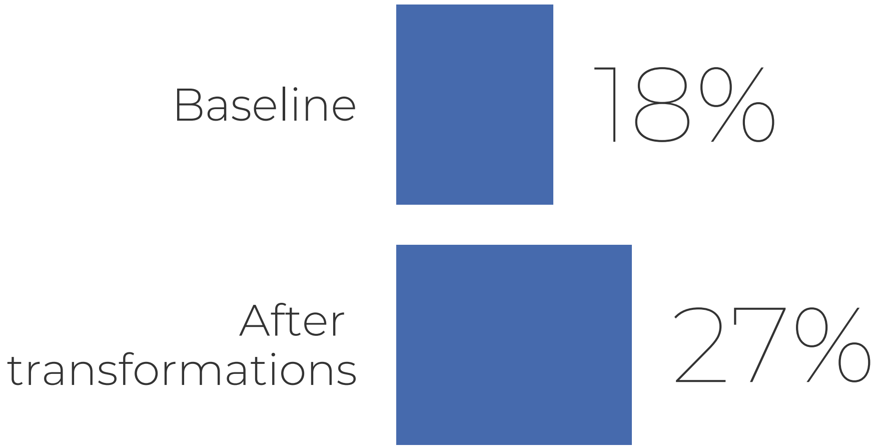


Area of Origin (WoE)

After apply feature engineering techniques, the fire incident model improved by 19 p.p

Model and Feature selection

- We used Decision Tree
- The model was evaluated using F1 score
- We kept features that contributed for 75% of the variance



There are many more opportunities to improve the Toronto Fire Incidents models

Next Steps

1. Understand how insurance companies assess risk related to fire incidents
2. Research the most common causes of exposure
3. Gather data for new potential features discovered in steps 1 and 2
4. Create pipelines based on the feature engineering analysis

Thoughts? Questions?

MADE WITH beautiful.ai