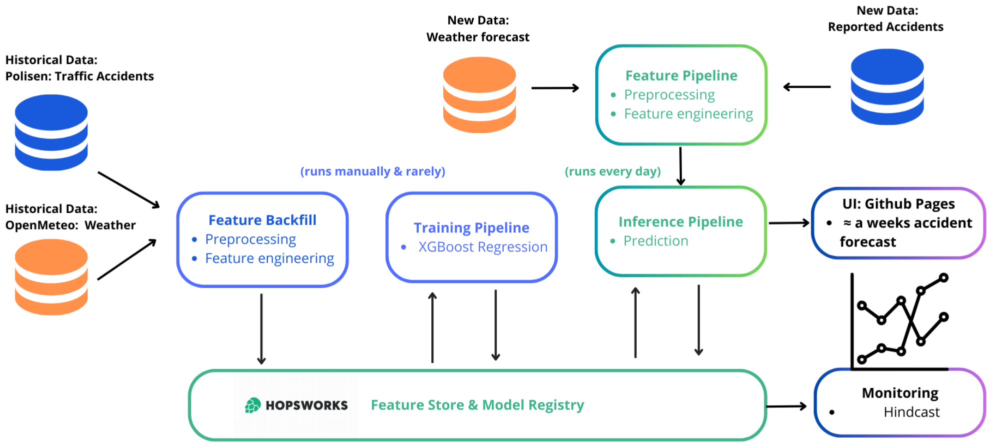
Real Time Traffic Accident Predictor

Saga Lundborg, Greta Jonsson
ID2223 HT24



Introduction

Should you decide not to take the car because the roads are currently looking like hockey rinks? Using combined data from Sveriges Polis and Weather forecasting from Open-Meteo we can model the risk of bad road conditions resulting in traffic accidents several days in advance.

APIs used

Polisen provides real time reports of traffic accidents.

The Open Meteo Weather Forcast gives us the ability to model accidents a few days in advance with the prior information of weather conditions.

Data

Target variable: Number of Accidents (Polisen)
This includes the following reported categories: Trafikolycka; Trafikolycka, personskada; Trafikolycka, singel; Trafikolycka, smitning från; Trafikolycka, vilt

Dependent variables (Open Meteo): Temperature 2m mean, Precipitation sum, Wind speed 10m max, Wind direction 10m dominant

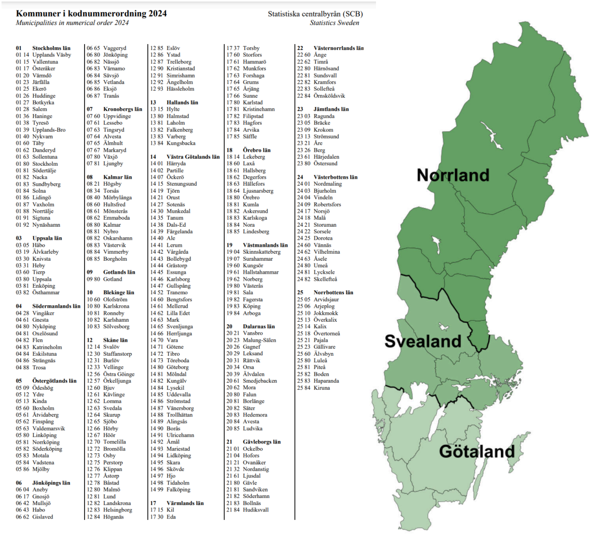
Additional dependent variables: Day of the week (Mon-Sun)

Model Choice

For this purpose of accident prediction a regression model is preferred over for example a LSTM model. Unlike stock prices, where the current price is heavily dependent on yesterday's price, the number of accidents could in some days, but must not depend on yesterday's data. One day there could be heavy rainfall causing more accidents and the next it will be sunny. Hence a regression model like XGBoost, being one of the top performing architectures for regression, is more suitable. It would be interesting to incorporate a hidden markov model in this prediction to incorporate transitions between different states over time, for example high precipitation to next day negative temperatures = high risk. However we did not complicate it and implemented it with XGBoost for this project.

Locational Filtering

Accident locations are reported in terms of Kommun. Due to the weather often being different from north to south, we are able to select different kommuner or groups of them formed by the counties (numbered 01-25). This filtering is applied on the historical data to train the model and the daily queries to the API. Due to national accidents summing up to around 20-30 each day, we recommend choosing larger areas of Sweden e.g. Götaland, Svealand and Norrland for the predictions since choosing only one county like Stockholm will result in very few accidents a day.



Feature engineering

We wanted to find out if days of the week were relevant features in the forecasting of accidents in addition to weather conditions since weekends can impact the sheer amount of movements on the roads.

Development Process

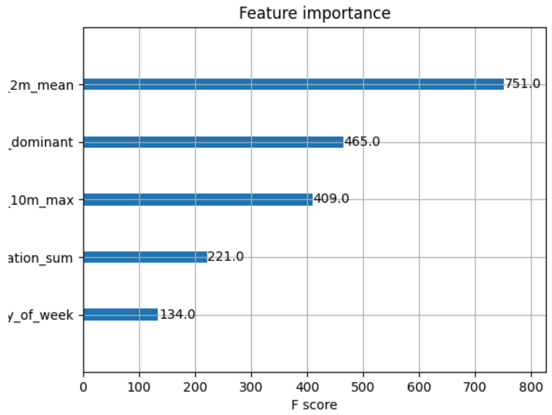
We first tried to use the API from the Swedish transport authority, Trafikverket, which was the original plan and had many useful data categories for example FronstDepthObservation, DeicingChemical (road salting) and reported accidents accessible through the API. The Trafikverket API also stated geographic location enabling use of separate models for separate parts of the country.

However, after some experimenting we found there was a limit on the historical data we could access, for example accidents only being shared for a day (Note: with special access given to collaborators of Trafikverket this method could be used to get the backfill with no expiry date on the data).

Hence we found historical data from another authority, Transportstyrelsen, with daily reports of accidents in the months December - January for the last 20 years. Unfortunately, when calling Trafikverket for their daily updates on accidents, they reported around 10 each day, but our Transportstyrelsen backfill stated 40-50 accidents a day. There was some asymmetry to the data, presumably reports filed (not shown on the API) dates after the accident happened.

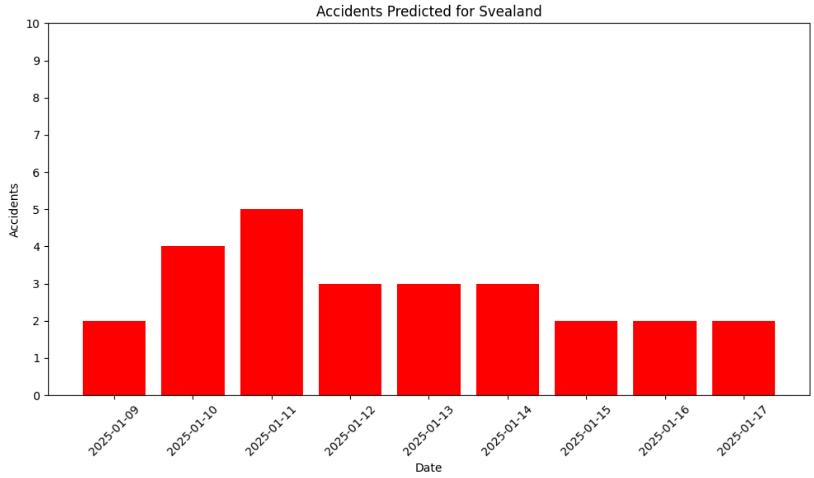
On our third attempt, we tried querying Polisens API, where we could access accidents reported back to 2024-08-01 (4 months) for a backfill and daily updates were possible.

Results

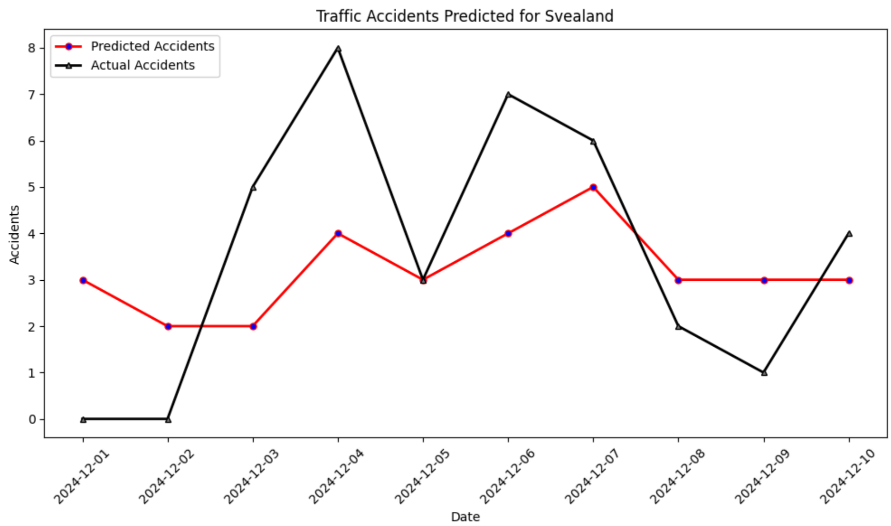


Feature Importance

Temperature 2m mean, Precipitation sum, Wind speed 10m max, Wind direction 10m dominant, Day of week



Date	Predicted Accidents
2025-01-09	2
2025-01-10	4
2025-01-11	5
2025-01-12	3
2025-01-13	3
2025-01-14	3
2025-01-15	2
2025-01-16	2
2025-01-17	2



Model Performance

MSE	6.0441537
R^2	-0.0976

Feature Importance
The most significant features influencing traffic accident predictions were weather variables. Temperature (2m mean), Precipitation sum, Wind speed (10m max), and Wind direction (10m dominant). These factors most correlate with accident risks, highlighting the role of weather in road safety. The day of the week influenced accident predictions, likely because traffic patterns and behavior change on different weekdays, like weekends.

Model Performance
The XGBoost regression model achieved a Mean Squared Error (MSE) = 6.044, reflecting the average squared deviation of predictions from observed accident counts. The R^2 = -0.098, indicating limitations in capturing the variance in accident numbers, possibly due to variability in data quality and underlying factors not included in the model.

Geographic Considerations
Using regional filters, such as Götaland, Svealand, and Norrland, helped improve the relevance of predictions across larger areas. However, when trying to make predictions for smaller regions, like specific counties like Stockholm, we had to expand our focus to include more areas due to the limited number of reported accidents.

Limitations
Inconsistencies in accident data from different sources, like Trafikverket and Transportstyrelsen, made it difficult to gather accurate information. As a result, we had to rely solely on police data, which limited the model's potential. Additionally, we did not include other variables, such as road salting levels or real-time traffic flow, which could have helped improve our predictions. The results show the value of using weather and time-based data to estimate accident risks. However, it also points to the need for more consistent and detailed datasets to boost the model's accuracy.

Discussion
The model shows promise, but it also highlights some key limitations. Using XGBoost for accident prediction made decent results, but the performance metrics indicate there's room for improvement. A major challenge was dealing with discrepancies in the data sources—Trafikverket's API provided limited historical data, while Polisens API offered more extensive backfill, though only up to a few months.

Feature engineering revealed that weather conditions, such as wind and temperature, played the most significant role, but adding temporal factors like the day of the week made sense as well. It's clear that accidents are influenced by a combination of environmental factors and human activities. We made the decision to keep the model relatively simple and stick with XGBoost, instead of exploring more complex approaches like HMM, given the project's limited scope. For future iterations, incorporating more advanced models or gaining better access to historical data (without expiry limits) could help improve predictions. After all, the accuracy of a model depends heavily on the quality of the data it's trained on.