Information Retrieval - Lab1 Report

1 Introduction

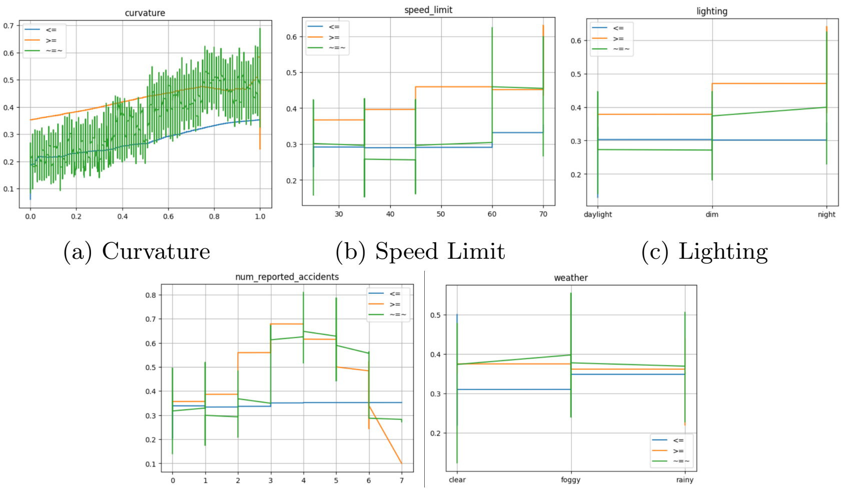
The paper looks at the Road Accident Risk dataset which can be found here:
https://www.kaggle.com/competitions/playground-series-s5e10

This is an artificially generated dataset, however the idea behind the exercise is to find patterns in data, and interpret the model's thought process, and for that purpose the fact that this is a toy problem only makes the relationships between parameters and target more apparent and easy to describe.

2 Data analysis

The dataset has 12 parameters, out of which 4 have shown a clear trend in the early data exploration process.

- curvature
- speed_limit
- lighting
- weather
- num_reported_accidents



(a) Curvature (b) Speed Limit (c) Lighting

(d) Reported Accidents (e) Weather

We can see some clear trends: As curvature increases, the probability of an accidents increases linearly with it. Similar trend can be noticed in relation to the speed limit. Poor lighting and foggy or rainy weather also increase the likelihood of an accident. One curious case is the fact that the model did not pick up any relationship between the time of day and target - that's very likely to be caused by the correlation between lighting and time of day - the model discarded the time of day as it was largely redundant. The last thing that is interesting is that as reported accidents grow the probability grows with them - however after the maximum at 4 and 5 accidents the probability drops - it's likely that some countermeasures were implemented to combat the riskiness of the road.

3 Explaining the model

The Variable Importance of out XGBoost model shows very similar trends to those noticed on the plots. Now to further our understanding of the model let's look at a few examples from the dataset which we inspected using Shap values.

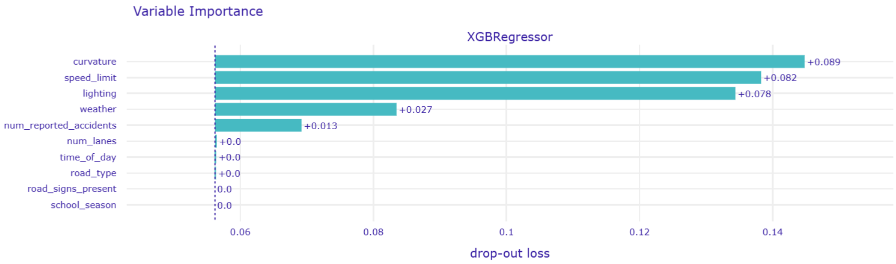


Figure 2: Variable Importance

3.1 Prediction p=0.13237005, y=0.13

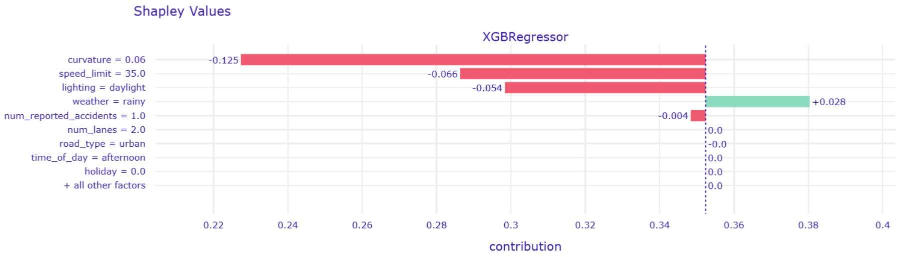


Figure 3: Variable Importance

Everything in this case points to the probability of accident being low - besides the rainy weather, which is the only parameter that contributes to increasing the prediction.

3.2 Prediction p=0.7318035, y=0.79

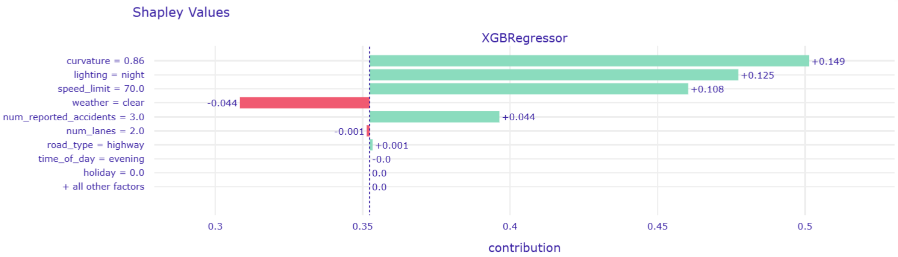


Figure 4: Variable Importance

Here the probability of an accident is high, because of the high speed, high curvature, lighting equaling night and number of reported accidents being equal to 3.

3.3 Prediction p=0.41356137, y = 0.47

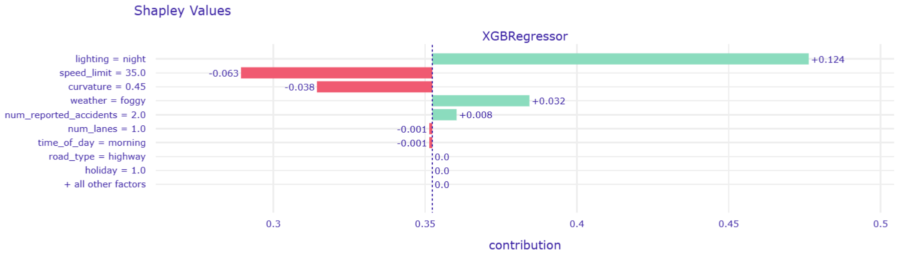


Figure 5: Variable Importance

In this case, there are things that both increase and decrease the riskiness of the road, such as low speed limit and low curvature. However, the fact that the lighting is bad and weather is foggy make the situation a bit more risky.

4 Takeaways

According to the model, the most effective ways to decrease road risk are to reduce curvature, lower speed limits, and improve poorly lit areas. Designing roads with gentler, longer curves instead of sharp turns would significantly decrease accident risk. Additionally, lowering speed limits—particularly on roads with a history of accidents—would make driving safer and help prevent future incidents.