Spatio-Temporal Road Failure Risk Prediction and Visual Evidence System for Denver

Rujaan Man Singh
Department of Data Science
Anderson College of Business and Computing
Regis University
3333 Regis Boulevard, Denver, CO 80221
rsingh001@regis.edu

March 2026

Abstract

The maintenance of urban infrastructure is commonly performed on a reactive basis, meaning that repairs are typically carried out only after damage has already occurred and been reported by residents. While public reporting systems such as 311 provide useful information regarding infrastructure issues, they do not assist cities in predicting where failures may occur in the future.

This practicum project develops a predictive system that identifies areas within Denver likely to experience road surface failures in the near future. Using historical Denver 311 service request data, a spatio-temporal machine learning pipeline was constructed to estimate failure risk at the grid-cell level.

The city was divided into spatial grid cells and complaint records were aggregated monthly in order to capture patterns of infrastructure deterioration over time. Several temporal features were engineered to represent recent failure activity. Multiple machine

learning algorithms were evaluated including Logistic Regression, Random Forest, and XGBoost. Among these models, XGBoost demonstrated the strongest predictive performance for structured tabular data and was therefore selected as the final predictive model.

To enhance interpretability, the system integrates Google Street View imagery that allows users to visually inspect road conditions associated with predicted high-risk locations. The final output is an interactive map that displays predicted risk levels together with street-level imagery. This project demonstrates how machine learning and geospatial analytics can support proactive infrastructure monitoring and improve prioritization of road maintenance activities.

1 Introduction

Maintaining road infrastructure is a major responsibility for municipal governments. Road surface failures such as potholes, cracks, and sinkholes can create safety hazards for drivers and pedestrians while also increasing long-term repair costs when not addressed early.

In many cities, infrastructure problems are primarily detected through citizen complaints submitted through public service systems such as 311. Although these systems provide valuable information, they operate in a reactive manner. Maintenance teams typically respond to issues after they have already occurred rather than identifying potential problem areas beforehand.

With the increasing availability of open municipal datasets, it is possible to apply data science techniques to analyze historical infrastructure records and identify patterns that may indicate future failures. Data-driven approaches are becoming increasingly important in urban planning and infrastructure management (**??**).

This practicum project focuses on developing a predictive model for road surface failure risk using Denver's publicly available 311 service request dataset. The project integrates several stages of the data science workflow including data preprocessing, spatial modeling, machine learning, and visualization.

2 Problem Statement

Municipal governments frequently rely on citizen complaints to identify road surface failures. While this mechanism helps detect existing infrastructure issues, it provides limited insight into where failures may occur in the future.

The main research question addressed in this project is:

Can historical 311 service request data be used to predict which areas of Denver are most likely to experience road surface failures in the following month?

To address this question, complaint records were transformed into a structured spatio-temporal dataset suitable for machine learning. Each spatial grid cell represents a location in the city and each monthly observation reflects recent failure activity in that location.

By predicting failure probabilities for each grid cell, the model enables identification of high-risk areas where inspections or preventative maintenance may be prioritized.

3 Methodology

This project follows several stages of the data science pipeline including data collection, preprocessing, feature engineering, machine learning modeling, and visualization.

The primary dataset used in this project is the Denver 311 service request dataset which contains citizen-reported infrastructure issues across the city. Complaint categories related to road failures were selected including potholes, potholes in alleys, sinkholes, and road hazards. Records with missing geographic coordinates were removed to ensure spatial accuracy.

To create a structured spatial framework, the city was divided into grid cells of approximately 300 meters. Each complaint record was assigned to a grid cell using geospatial join operations. This grid-based representation converts irregular complaint locations into consistent spatial units suitable for predictive modeling.

The dataset was aggregated monthly to construct a spatio-temporal panel dataset. Feature engineering was performed to capture recent failure activity including counts of failures in the previous one month, three months, six months, and twelve months.

Several machine learning algorithms were evaluated including Logistic Regression, Ran-

dom Forest, and XGBoost. Logistic Regression served as a baseline model, while Random Forest provided a nonlinear ensemble approach. XGBoost was selected as the final model because of its ability to capture complex relationships within structured data (**?**).

The modeling pipeline was implemented using Python and the Scikit-learn library for machine learning and data analysis (**?**).

4 Data Analysis and Visualization

Exploratory data analysis was conducted to examine spatial and temporal patterns within the Denver 311 dataset. Initial visualizations revealed that complaints tend to cluster in specific regions of the city rather than being evenly distributed.

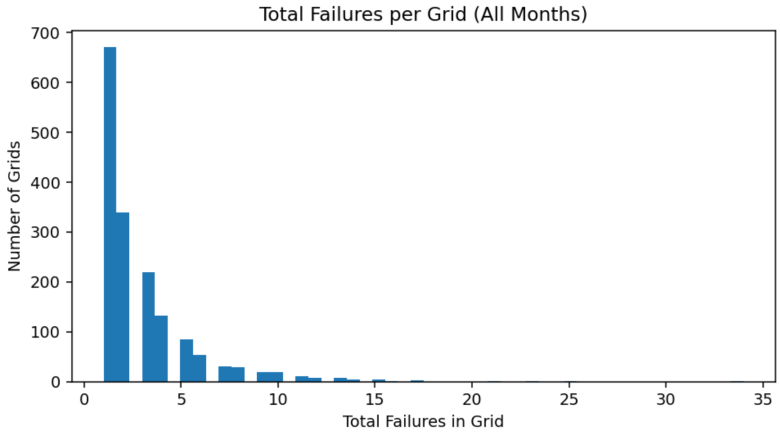

Figure 1: Distribution of road failure events across grid cells.

The model produces predicted risk scores for each grid cell. The distribution shows that most grid cells have relatively low predicted risk while a smaller number of areas show higher probabilities of road failure.

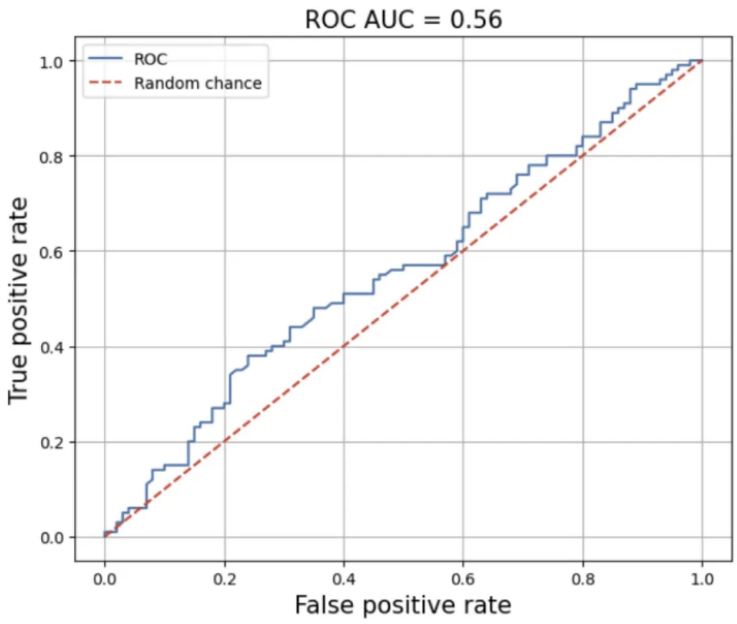

Figure 2: ROC curve illustrating predictive performance of the XGBoost model.

To further understand the spatial distribution of predicted risk, a grid-based risk map was generated.

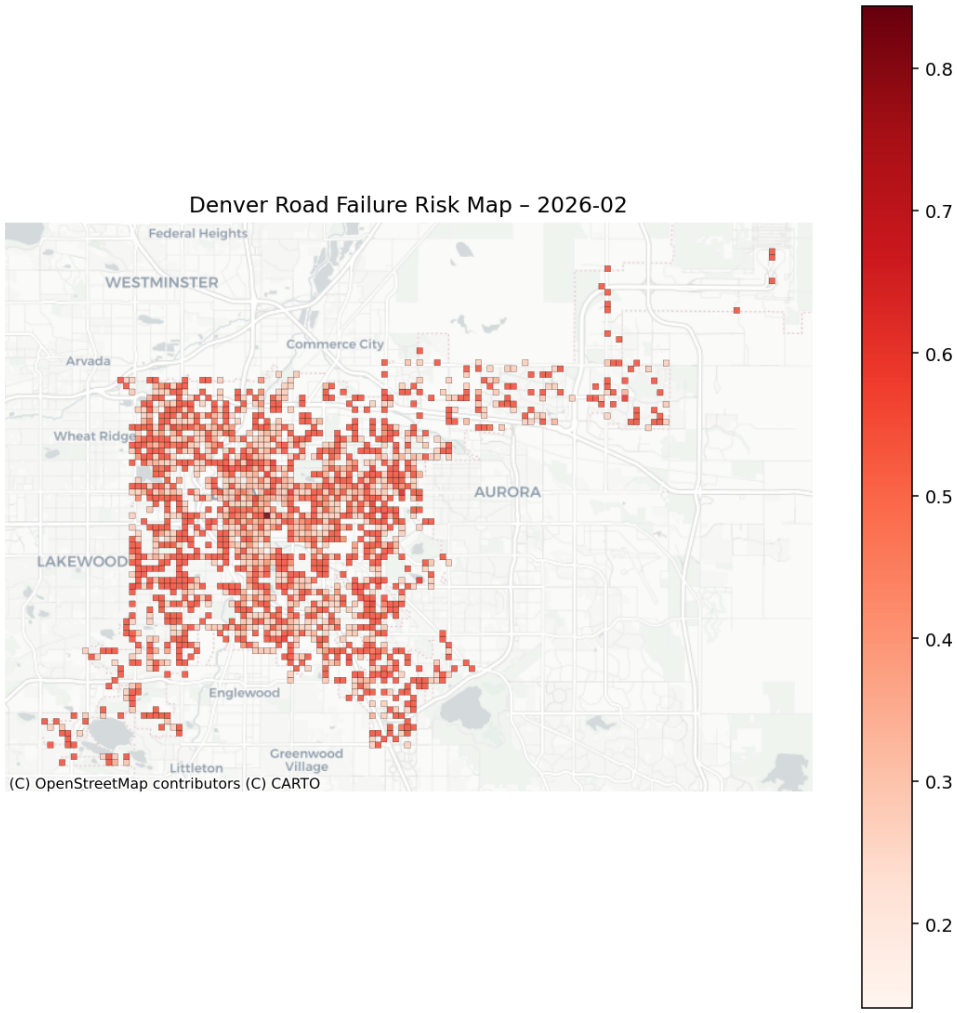

Figure 3: Spatial risk map showing predicted road failure probabilities across Denver.

To improve interpretability, an interactive visualization was developed using the Folium library. Users can explore predicted risk levels across the city and view Street View imagery associated with each grid cell.

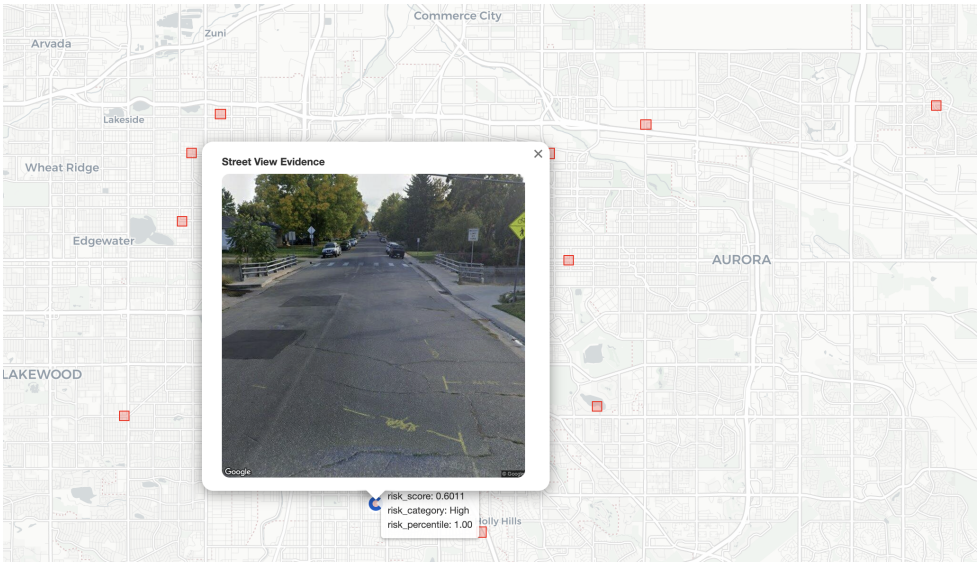

Figure 4: Interactive map displaying predicted risk levels and associated Street View imagery.

5 Expected Outcomes

The primary outcome of this project is a machine learning model capable of estimating the probability of road surface failures at the grid-cell level. These predicted probabilities serve as risk scores that allow areas of the city to be ranked according to their likelihood of experiencing infrastructure problems.

Another significant outcome is the development of an interactive visualization system that combines predictive analytics with street-level imagery. This interface allows users to explore high-risk areas and visually inspect road conditions associated with those locations.

6 Timeline

The practicum project was completed in several stages. The initial phase involved reviewing relevant literature and identifying suitable datasets. The next stage focused on data collection, cleaning, and exploratory data analysis.

Following this, spatial grid construction and feature engineering were performed to convert complaint data into a structured spatio-temporal dataset. Machine learning models were then developed and evaluated.

The final phase involved integrating Street View imagery, developing the interactive visualization system, and preparing the final report and presentation.

7 Conclusion

This project demonstrates how data science techniques can be applied to real-world urban infrastructure challenges. By analyzing historical 311 complaint data, patterns can be identified that help predict where road failures are likely to occur.

The project integrates geospatial data processing, machine learning modeling, and data visualization into a unified predictive framework. The inclusion of street-level imagery provides visual context that improves interpretability and allows users to validate model predictions more effectively.

While the current model relies primarily on historical complaint data, future work could incorporate additional information such as traffic patterns, weather conditions, or pavement age to improve predictive performance.

References

Anselin, L. (1988). *Spatial econometrics: Methods and models*. Springer.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794).

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning*. Springer.

Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*.

Piryonesi, S. M., & El-Diraby, T. E. (2021). Using machine learning to examine the impact of type of performance indicator on flexible pavement deterioration modeling. *Journal of Infrastructure Systems, 27*(1). https://doi.org/10.1061/(ASCE)IS.1943-555X.0000602

Fathi, A., Mazari, M., Saghafi, M., Hosseini, A., & Kumar, S. (2019). Parametric study of pavement deterioration using machine learning algorithms. In *International Airfield and Highway Pavements Conference 2019* (pp. 37–47). American Society of Civil Engineers (ASCE). https://doi.org/10.1061/9780784482476.004

Denver Open Data Portal. (2024). Denver 311 service requests dataset.