Imperial College London

Using Machine Learning Algorithms to Predict and Visualise Flood Risk in the UK

Authors:

Benjamin Duncan

Junyi Li

Wenxin Li

Zehong Lin

Nadia Mason

Ziheng Meng

Georgia Ray

Emilio Tinoco Robledo

Enzo Zhang

Date: November 24, 2023

GitHub Repository: ads-deluge-nene (click here)

Introduction & Motivation:

Due to climate change, flooding is becoming an increasingly important issue. This project uses Machine Learning Algorithms to predict flood risks in the UK, using location coordinates, soil type, elevation, & population and pet data. Our goal is to develop a predictive tool that classifies flood probabilities and the risk value defined by:

$$R := 0.05 \times (\text{total property value}) \times (\text{flood probability})$$

where the approximate flood probability that there is one event in a given year for the ten risk classes can be assumed to be:

Class	Flood event probability
10	5%
9	4%
8	3%
7	2%
6	1.5%
5	1%
4	0.5%
3	0.3%
2	0.2%
1	0.1%

Table 1: Flood Event Probability by Class

Methods

RiskLabel: Initially our evaluation of SVM and GradientBoostingClassifier for predicting location-based risks was hindered by imbalanced data, preventing generalisation. We also found that converting risk labels to probability for a regression approach did not improve performance. Therefore we proceeded with classification. This led us to shift to KNeighborsClassifier and RandomForestClassifier, which excelled in managing data clustered by local authorities. Feature selection was conducted via permutation importance test, with 'easting' and 'northing' as key predictors.

Median House Price: Our optimised model uses a RandomForestRegressor with feature scaling via StandardScaler, fitting on coordinates, population averages, and mean household pets. We excluded soilType and elevation features due to their negligible impact on performance relative to increased model complexity. For unseen postcodes, it assigned regional mean feature values, aiming to improve generalisation - however this may remain a point for furture exploration. Alternative SVR and KNN models were dismissed due to their sensitivity to dataset outliers and the intensive computation required for SVR's kernel and hyperparameter tuning.

Historical Flooding: The final historical flood model employs a ColumnTransformer with a SimpleImputer for both numerical (mean strategy) and categorical (most frequent) data, followed by RobustScaler for numerical data and OneHotEncoding for categories. Future improvements could involve using a geographically correlated "local median" for imputation to potentially enhance model performance. A RandomForestClassifier is then used on all features.

Feature Importance:

Exploratory Data Analysis revealed that geographic proximity is a key determinant, with coordinates being the most critical predictors. The Median House Price model improved significantly with additional features like average population and household pets and therefore includes more features. Historical Flooding model also maintained performance by using all features.

Visualisation Demonstration:

For interacting with the default datasets and the predictions of the annual flood risk, median house price and the risk label (or flood class), three functions are available by importing the `flood_tool` package. The first function {`display_features:`} allows the user to interact with both the default dataset and the predictions. It is strongly recommended to review the documentation of the function carefully to ensure the proper parameters have been selected. There are two different ways of using the {`display_features:`}. The user can pass a path of a file already stored in the resource directory and pass a data frame with the predictions.

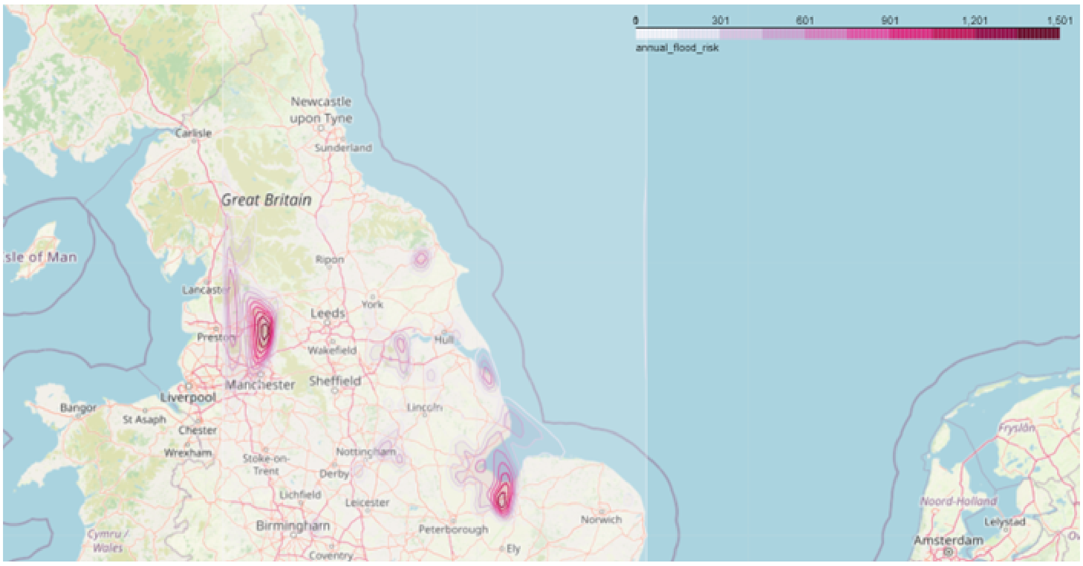


Figure 1: Annual Flood Risk prediction. The contour lines represent zones with similar risk values. The units are current GBP.

The function {`add_countours:`} can be called separately to display individual contour lines from datasets such as `wet_day.csv` and `typical_day.csv`. The user should select whether the parameter plotted is a rainfall characteristic or a river property. The map displayed is interactive and layers can be turned off or on.

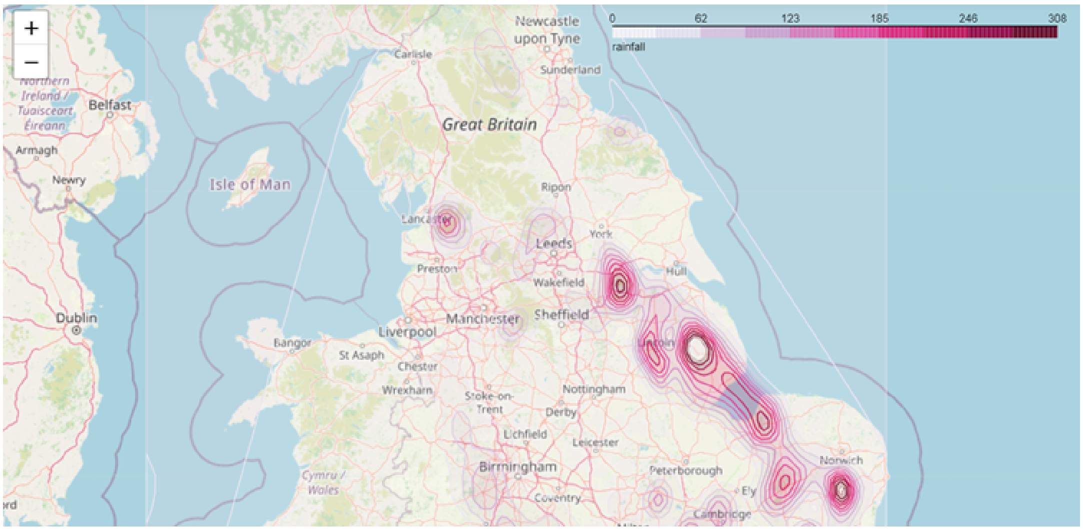
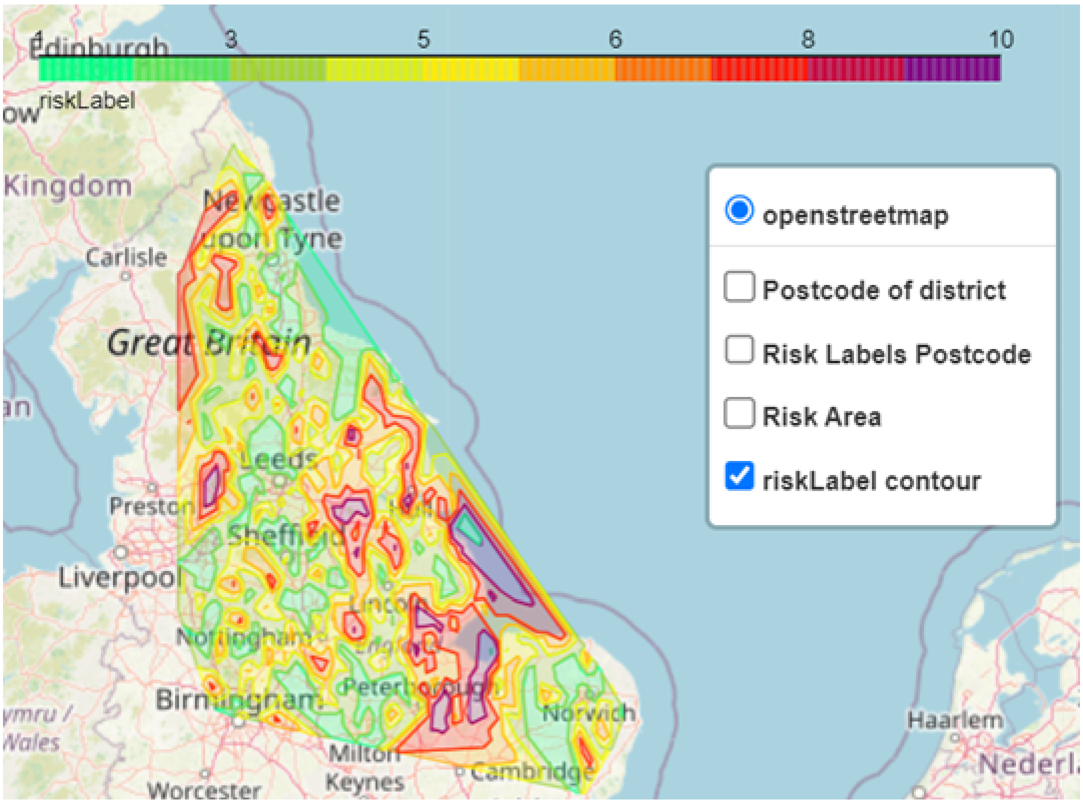


Figure 2: Wet day contour lines represent zones with similar rainfall values in mm. Some of the spots highlight the zones with high risk.



● openstreetmap

□ Postcode of district

□ Risk Labels Postcode

□ Risk Area

☑ riskLabel contour

Figure 3: Contour lines colored by the risk class or risk label. The higher risk class match moderate the zones with high annual flood risk and zones with extreme values of rainfall