D682 Task 2

Dennis J Pothier

Student ID: 011690596

27Jul2025

D. Narrative Report

1. Description of Optimization Techniques

SelectKBest was employed to reduce the input features and make the model more efficient and understandable. Most environmental health data sets contain many variables, but there are not many that significantly contribute to predictions. SelectKBest employs F-regression scoring ranking and chooses the features with maximum statistical correlation with the target variable. In return, this reduces the risk of overfitting and makes explaining the model easier for stakeholders by focusing on the most significant environmental indicators (Guyon & Elisseeff, 2003).

RandomizedSearchCV was employed for hyperparameter optimization because it trades between completeness and computational cost. In contrast to GridSearchCV, which checks all combinations in detail, RandomizedSearchCV makes random draws of specific parameter settings from a distribution, making it computationally efficient and scalable. This is particularly true with computationally intensive models like Random Forests (Bergstra & Bengio, 2012).

2. Regularization Techniques Overview

Ridge Regression was selected as a baseline model employing L2 regularization to mitigate multicollinearity. Environmental data generally have highly collinear features like temperature and humidity. Ridge regression penalizes significant coefficients such that a feature cannot take control of the model, thus avoiding overfitting and improving model stability (Hoerl & Kennard, 1970).

Random Forest parameters were regularly controlled by restricting tree depth, minimum split, and leaf sample sizes. These modifications prevent trees from becoming overly complex, a well-documented flaw of ensemble methods like Random Forests. Structural regularization improves generalization and consistency across subsets of data (Louppe, 2014).

3. Explanation of Ensemble Learning Techniques

Voting Ensemble was used to combine predictions from different types of models (Random Forest, Ridge Regression, Gradient Boosting) to leverage their complementary strengths. It reduces model variance and enhances robustness by voting the output of different algorithms, which is useful where no single model works best on all occasions (Opitz & Maclin, 1999).

XGBoost was selected as it works better with structured data and includes in-built regularization properties and missing value handling. It can model high-order nonlinear interactions between environmental variables and includes built-in facilities for avoiding overfitting without extensive hand-tuning of parameters (Chen & Guestrin, 2016).

4. Description of Evaluation Metrics

Root Mean Squared Error (RMSE) was employed since it measures prediction accuracy in the same units as the target variable and heavily penalizes high errors. This is important in health applications, where high prediction errors could result in under- or over-interventions. RMSE is also widely known and employed across regression analysis (Chai & Draxler, 2014).

R-squared (R²) was used to measure the amount of variance in the health risk scores explained by the model. It gives a simple, easy-to-understand measure of model fit and allows for comparison across differing model approaches, enabling stakeholders to appreciate the value added by the predictive system. It is particularly valuable when communicating with non-technical stakeholders. (Chai & Draxler, 2014)

5. Overall Business Impact and Performance Analysis

The full optimized model had measurable performance improvements over the baseline Random Forest model. Test RMSE dropped from 0.1146 to 0.1079, a 5.8% improvement, whereas R² increased from 0.9711 to 0.9743. Cross-validation metrics were similarly in the same direction, with an improvement of 14.23% in RMSE and over a 1% increase in R².

These improvements imply that the optimized model is not only more precise but also more generalizable. For this business case, predicting environmental health risk translates into better early warnings, more targeted interventions, and stronger policy decisions. Making strong, interpretable predictions enables public health agencies and environmental regulators to target resource investments and decrease risk more efficiently.

E. References

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*, 281–305. <https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf>

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development, 7*(1), 1247–1250. <https://gmd.copernicus.org/articles/7/1247/2014/>

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery. <https://arxiv.org/abs/1603.02754>

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182. <https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf>

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. <https://doi.org/10.2307/1267351>

Louppe, G. (2014). Understanding random forests: From theory to practice (Doctoral dissertation, University of Liège). *arXiv*. <https://arxiv.org/abs/1407.7502>

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research, 11*, 169–198. <https://www.jair.org/index.php/jair/article/view/10302>