# Variable Selection

We are dealing with a rather small data set with about 25 rows for 7 input variables. Yet, it is good practise to build statistical and machine learning models with a set of input variables which are statistically significant.

Based on the EDA, we figured that 3 variables are highly correlated – creating the problem of multi-collinearity. The variable selection stage will also help us deal with this issue. For this reason, we explore two methods for variable selection – (1) Lasso Regression and, (2) Random forest.

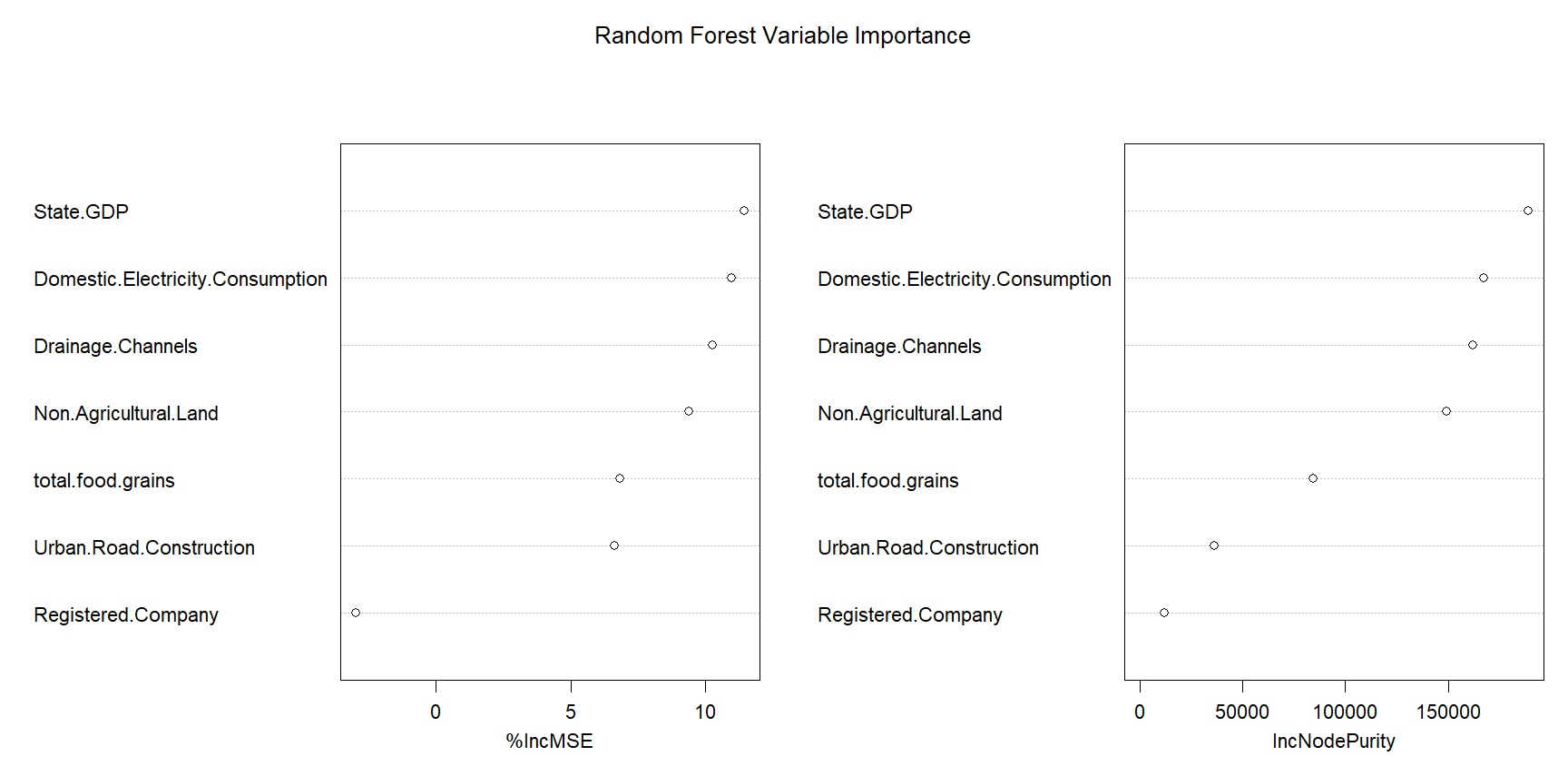
## Lasso Regression

Lasso regression will zero-out the coefficient of irrelevant variables thus eliminating them. The issue of multicollinearity will be dealt via regularization.

For lasso, we first need to find the most optimal value of the regularisation parameter lambda. For this we perform cross validation and find lambda = 2.05 minimizes MSE.

|  |  |  |
| --- | --- | --- |
| **Selected Variable** | **Coefficient** | **Interpretation** |
| Intercept | -15.67 | Baseline flood count when all predictors are at their normalized minimum |
| Total food grains | -187.93 | Higher food grain productivity (yield/hectare) is negatively associated with floods — possibly capturing a “rural buffer” or efficient land use |
| Domestic Electricity Consumption | +869.23 | Very strong positive association — suggests urban housing growth → more floods |
| Urban Road Construction | +114.58 | More roads → more impervious surface → increases flood risk |
| Registered Company | -302.64 | Surprisingly negative — may reflect formal commercial growth occurs after infrastructure development, not directly driving floods |
| Drainage Channels | -4.77 | As expected, drainage has a slight negative impact on flood count, but the effect is very small |

## Random Forest



The Random Forest model (with a 83% var explained, indicating a very good fit) gives us a list of important variables based on two criteria's - % increase in MSE and Increase in node purity.

|  |  |  |
| --- | --- | --- |
| **Selected Variable** | **% Increase in MSE** | **Notes** |
| State GDP | 11.44 | Highest impact — confirms macro-growth–flood linkage |
| Electricity | 10.97 | Very strong predictor again — validates urban pressure |
| Drainage Channels | 10.26 | RF considers it very useful (though counterintuitive direction) |
| Non Agricultural Land | 9.40 | Again ranks high — land use conversion impact confirmed |
| Food Grains / Roads | ~6.7 | Still relevant, but not dominant |
| Registered Company | −2.98 | Negatively associated or just noise |

### Which makes more sense?

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Random Forest** | **Lasso** |
| **Handles non-linearity** | ✅ Yes | ❌ No |
| **Deals with multicollinearity** | ✅ Strong (via tree structure) | ✅ Medium (via regularization, but may drop good variables) |
| **Interpretability (coefficients)** | ✅ with XAI techniques | ✅ Easier |
| **Variable stability** | ✅ Good | ❌ Can be unstable with correlated vars |
| **Better performance** | ✅ 82.3% variance explained | Lasso doesn’t provide R² directly, but likely lower |
| **Penalty awareness** | ❌ None (fully greedy) | ✅ Tunable via λ |

**Multicollinearity Handling**

* **Lasso** dropped State GDP because it is highly correlated with Electricity (r ≈ 0.98) and Drainage.
* **Random Forest** retained all three and still gave them importance based on actual predictive split contribution.

So, **Random Forest handles multicollinearity better** *in terms of preserving strong features even if correlated*.

### Final Choice

Instead of choosing Food Grains, we select Urban Road because they capture the essence of how cutting mountains and disrupting natural cover to build infrastructure affects natural calamities like flood.

* Roads fragment landscapes, accelerate runoff, block natural water flows.
* Construction often involves deforestation, hill-cutting, and drainage disturbance.
* Floods in Guwahati, for example, have often been linked directly to unplanned roadwork and urban encroachments.

On the other hand, food grains are more indirect:

* Agricultural expansion can reduce permeability, but so can food-efficient farming methods with less sprawl.
* It’s harder to distinguish whether productivity gains came from **intensity** (irrigation) vs **expansion** (more land use).

Additionally, the difference in % increase in MSE between Food Grains and Urban roads – is negligible. It makes sense to go with a variable that has stronger contextual significance.

|  |  |
| --- | --- |
| **Final Selected Variable** | **Justification** |
| State GDP | Broad proxy for macroeconomic activity, construction booms, and capital-intensive growth |
| Electricity Consumption | Captures urban household expansion and increasing residential density |
| Drainage Channels | Infrastructure response — may reveal reactive vs proactive investment effects |
| Non-Agricultural Land | Signals land-use change, particularly conversion to built-up, commercial or residential use |
| Urban Road Construction | Reflects impervious infrastructure growth, road widening, hillside cuts — direct physical disruptors of natural flow |