**Capstone Outcome**

**Objective**:

The primary objective is to devise an explainable model that can effectively rank RCs based on their needs of investment.

**Methodological Considerations:**

1. **Principal component analysis**

Guided by mentorship and outcomes from the correlation analysis, we pinpointed the most suitable variable groups for Principal Component Analysis. Here's the process:

* Step 1: Identify the 20 most correlated variables for each of the 64 variables in the dataset.
* Step 2: Keep only variable pairs with a correlation above 0.4, and group these highly correlated variables.
* Step 3: Normalize the data, perform singular value decomposition on each identified variable group, and derive the corresponding PCA factors.

The generated PCA factors are treated as eleven indexes representing different domains with respect to flood: damage index, government investment index, inundation index, rain index, river index, infrastructure population index, road index, terrain index, land index, drainage index, electricity index. In our following steps, we are going to utilize these indices as the dataset for developing classification models. Detailed information of these indices is appended at the end of the note.

1. **K-means** **Clustering (Finished)** :

While clustering effectively groups similar entities, it doesn't inherently offer a ranking system, which aligns with our ultimate objective. To establish this ranking within the clustered groups, we sum ['government\_investment\_index', 'damage\_index', 'inundation\_index', 'rain\_index']. Instead of assigning a precise ranking to each Revenue Circle (RC) from 1 to 180, we employed K-means clustering to categorize them into 6 groups, ranging from 1 to 6. This approach is deemed appropriate as it is challenging to distinctly prioritize RCs when they exhibit similar levels of damage and characteristics in other aspects. The resulting classification is justifiable as the flood index aptly mirrors the extent of damage, inundation levels, and rainfall in specific areas. Furthermore, it reveals that previous tender allocations were not solely based on damage levels, as some of the most damaged areas sometimes don’t receive tenders.

1. **Random Forest (Finished)** :

In harnessing the robustness of the Random Forest classification model, we began our analysis of flood damage by applying it to the PCA factors. We take the damage index as the target variable and use the remaining indices as independent variables. The strength of Random Forest lies in its ability to handle a large number of input variables and its effectiveness in classifying complex datasets, making it particularly suitable for our analysis involving diverse factors. This model excels in reducing overfitting, providing more reliable predictions compared to simpler models.

Consultations with our mentors highlighted that variables like total government awards and cumulative rainfall have a sustained influence on the effects of flooding. This realization underscored the need to incorporate their cumulative values into our analysis, ensuring a more comprehensive understanding of the long-term impacts of these factors. Such insights were pivotal in shaping our approach, allowing for a more nuanced and effective assessment of flood vulnerability across the Revenue Circles. Consequently, we conducted PCA on both the original dataset and a modified dataset that incorporated cumulative values for certain variables, allowing us to compare the performance of the models.

Result:

A screenshot of a table

Description automatically generated

1. **Linear Regression (terminated)** :

Limitation: We cannot address issues related to non-linear relationships and the presence of zero inflation in the data. Data transformation doesn’t work, Ridge / Lasso also doesn’t work. The data is also seriously zero-inflated.

1. **AHP Model (terminated)** :

Limitation: Traditional AHP relies heavily on expert judgment for weighting, which is not accessible to us, and it also doesn't accommodate the high number of variables (maximum of 9) in our dataset effectively.

1. **Non-linear Regression (terminated)** :

Limitation: Difficulties arise with feature selection and the interpretation of models, especially when considering interaction terms. Moreover, many interaction variables are not reasonable.

1. **Logistic Regression (terminated)** :

Limitation: Logistic regression is typically not suitable for a continuous outcome like the number of people affected.

Potential Adaptation: By categorizing the population affected into a binary variable (e.g., 1 for affected population > 10,000; 0 otherwise), logistic regression could assess the probability of a significant impact on the population. However, this deviates from the main objective, which is to understand the intensity, not just the occurrence, of the impact.

1. **Poisson Regression (terminated)** :

Limitation: Poisson regression is more apt for count data representing the number of occurrences within a fixed period or space rather than the continuous scale of the population affected.

Potential Adaptation: Could be considered if the goal shifts towards analyzing the frequency of floods or disasters rather than the impact scale.

**Possible data manipulation:**

We explored possible cumulation for all continuous variables, and found these variables are meaningful:

**Cumulate over every two month**: ['Embankments affected', 'sum\_rain', 'flooded\_vegetation']

**Cumulate over every three month**: [ 'sum\_rain']

We also transformed all continuous variables to categorical variables (high, mid, low), and found these variables are meaningful:

**Meaningful categorical**: ['max\_rain', 'mean\_rain', 'sum\_rain', 'inundation\_pct', 'inundation\_intensity\_mean', 'riverlevel\_mean', 'riverlevel\_min', 'riverlevel\_max', 'sum\_aged\_population', 'sum\_young\_population', 'sum\_population', 'schools\_count', 'health\_centres\_count', 'rc\_piped\_hhds\_pct', 'drainage\_density', 'old\_ratio', 'young\_ratio']

Old ratio is calculated by (old\_population / total population), young ratio is calculated by (young\_population / total population).

**However, because of the limitation of PCA, we didn’t apply any categorical transformations. Our analysis revealed that incorporating cumulative variable effects did not yield significant results. Consequently, we decided to omit these meaningful cumulative variables from our final dataset. These cumulative variables might be meaningful for some other model.**

**Index for our random forest model and K-means model:**

**government\_investment\_index**: [total\_tender\_awarded\_value, Preparedness Measures\_tenders\_awarded\_value, Immediate Measures\_tenders\_awarded\_value];

**damage\_index**: [Total\_Animal\_Affected, Population\_affected\_Total, Crop\_Area, Relief\_Camp\_inmates, Human\_Live\_Lost, Roads, Bridge];

**inundation\_index**: [inundation\_pct, inundation\_intensity\_mean, inundation\_intensity\_mean\_nonzero, inundation\_intensity\_sum];

**rain\_index**: [max\_rain, mean\_rain, sum\_rain];

**river\_index**: [riverlevel\_mean, riverlevel\_min, riverlevel\_max];

**infrastructure\_population\_index**: [health\_centres\_count, schools\_count, sum\_aged\_population, sum\_young\_population, sum\_population, net\_sown\_area\_in\_hac];

**road\_index**: [rail\_length, rail\_count, road\_length, road\_count];

**terrain\_index**: [slope\_mean, elevation\_mean, distance\_from\_river, mean\_ndvi, trees];

**land\_index**: [bare\_ground, water, rangeland, mean\_cn];

**drainage\_index**: [drainage\_density, crops];

**electricity\_index**: [avg\_electricity, rc\_piped\_hhds\_pct];

**Dropped variables**: [SOPD\_tenders\_awarded\_value, SDRF\_tenders\_awarded\_value, RIDF\_tenders\_awarded\_value, LTIF\_tenders\_awarded\_value, CIDF\_tenders\_awarded\_value, Others\_tenders\_awarded\_value, Total\_Animal\_Washed\_Away, Total\_House\_Fully\_Damaged, Embankments affected, Embankment breached, ndbi\_mean, mean\_sexratio, flooded\_vegetation, built\_area, clouds, avg\_tele, rc\_nosanitation\_hhds\_pct]