Bag Tracking and Alert System

• • •

Ministry Category: Department of Posts, Ministry of Communications

Problem Code: #POS1

College Code: 1-2811634924

Team Name: Connected

Team Leader: Soumyadeep Roy

Solution Approach

Step 1 :

We assume logs, booking and delivery details as visible during Speed Post tracking will be provided.

1. Date, Time, Office and Event
2. Consignment Number
3. Booking Details : *Booked On, Booked At, Category, Tariff , Article Type and Destination Pin-code*
4. Delivery Details : *Status, Delivered to and Delivered on*

Based on previous studies on postal system, the following article details are needed *:*

Volume, Weight and Address method(Handwritten or typed)

Step 2 :

Derive post office-centric information from individual transactions by suitable data manipulation. Perform feature-engineering to derive more meaningful attributes.

Visualise these meaningful attributes to get an understanding of the overall distribution of traffic and misplacement. These visualisations will be reported only to the Postal department and not to the general public.

Misclassification Intensity = No. of misclassifications in last 6 months

Traffic Intensity = No. of items received + No. of items dispatched

Step 3 :

Use datasets of causal factors from data.gov.in. Use provided Data API where possible. The factors were intuitively selected where we tried to bring together all attributes that can contribute to delays and losses.

1. Geotagged Post office information : Post offices all over India with their latitude and longitude coordinates are provided. Use linear regression model to estimate node-node delivery time, to make a more significant and meaningful approach. Based on intuition,

Estimated node-node delivery time = f(Euclidean distance , Article Type)

2. Terrain : Assign each district to one of the 3 categories as the terrain may make a significant difference.

Categories : easy(1), moderate(2), difficult(3)

3. Beneficiary count per post office : Equal to the number of Voter Card Holders. Use the data to identify the Post office node type.

Node types : link, end

Step 4 :

Assign each office to one zone, in our case district-wise zoning is done.

Reason : Reducing the number of nodes, by clustering based on geographical factors.

Add attributes corresponding to the causal factors for each record. The attribute values will be based on their respective zones. All offices belonging to the same zone will have the same causal factor attribute values

We now have an informative and contextual dataset to start modelling

Step 5 :

The new intuitive dataset is clustered using *Bradley-Fayyad-Reina(BFR) algorithm*. It is a variant of k-means clustering algorithm designed for very large datasets. Initially we will use k-means clustering function of Spark MLlib .

Clustering is done based **only** on postal system behaviour w.r.t delays and losses and their causal attributes. A maximum of 4 clusters will be allowed.

Reason : Purpose is to categorise them instead of rating them.

Step 6 :

These clustering results are effectively visualized and corresponding statistics are properly documented. These reports are sent to the Postal department only and is not for general public.

Reason for clustering :

1. Previously our results were zone-specific. Clustering disaggregates the training set based on this result.
2. Have strong influence over the **decision-making process** of administrators and policy-makers, since they now have knowledge and facts at their disposal.

Step 7 :

Each cluster will be separately modelled by *Least-squares regression analysis*. For each category, significant factors obtained will be duly reported to the Postal department.

Step 8 :

These regression models will be used for estimating Misclassification Intensity zone-wise as well as for any specific post office.

Significant factors behind the delays and losses are also known, which will help us in making effective **recommendations for improvement.**

Step 9:

The estimated intensity and cost associated with the losses and delays , is compared with the real life data, in order to evaluate our model. Accident risk is computed for each zone.

*Accident risk = x*Estimated value + y*Observed value*

where x , y solely depends on model accuracy

This Accident Risk Estimation model will **evolve with time** as the insights gained will be used to improve our choice of : *Causal factors, features, machine learning algorithms*

Step 10:

The zones are ranked based on the estimated accident risk. These results are visualized, highlighting the significant causal factors. The results will be reported to the general public through this web application. Suggestions will be duly reported to the Postal department regarding:

1. Improving the **Quality of Service** *(Speed of delivery, reliability, customer satisfaction)* by studying usage trends.
2. Reducing the number of delays and losses by utilizing the evolving Accident Risk Estimation model.

Step 11:

A Recommender service, created using the results from the Accident Risk estimation model, will be provided for general public usage.

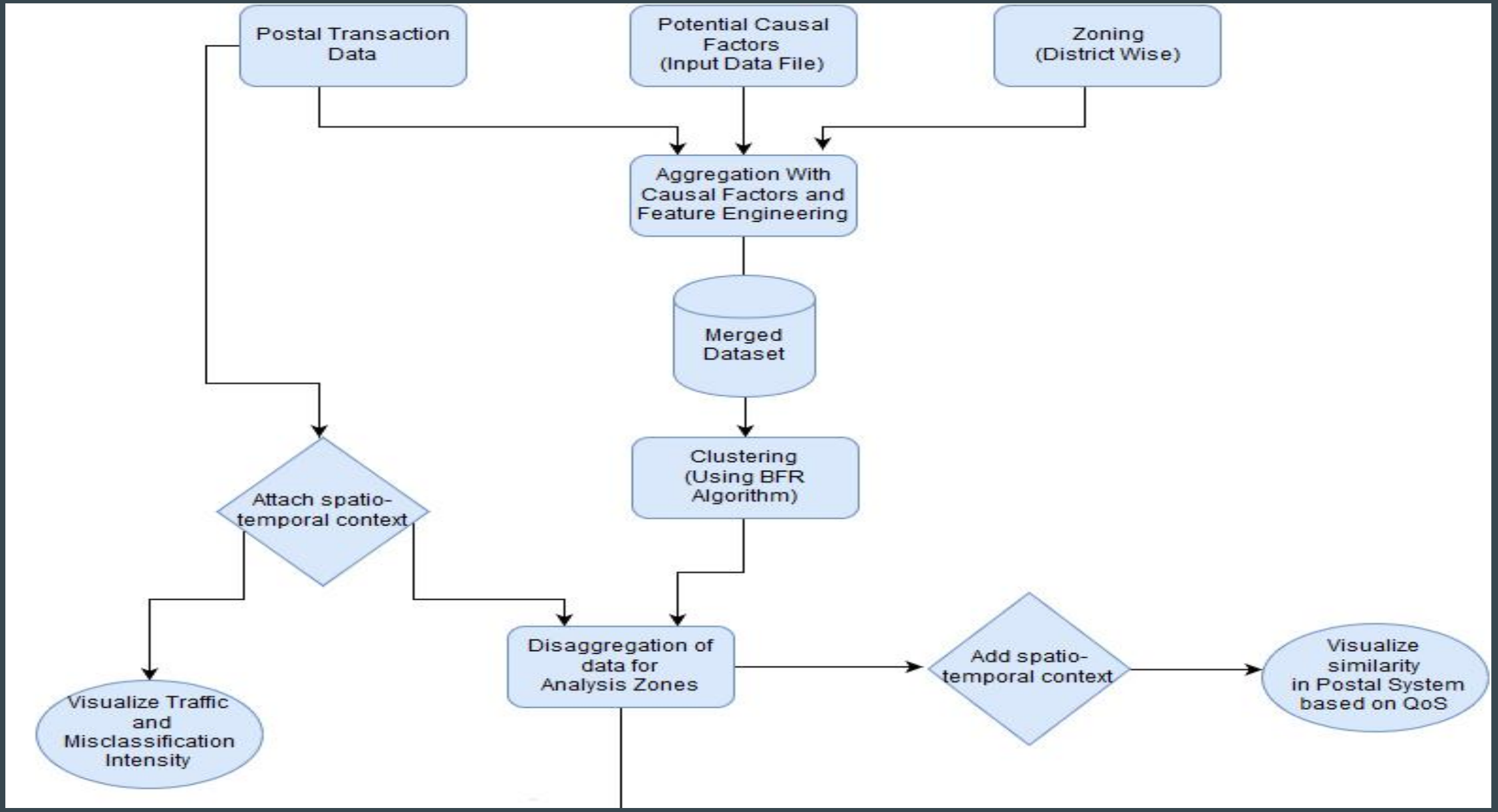
Input : Location(L), Postal facility type(F) from user

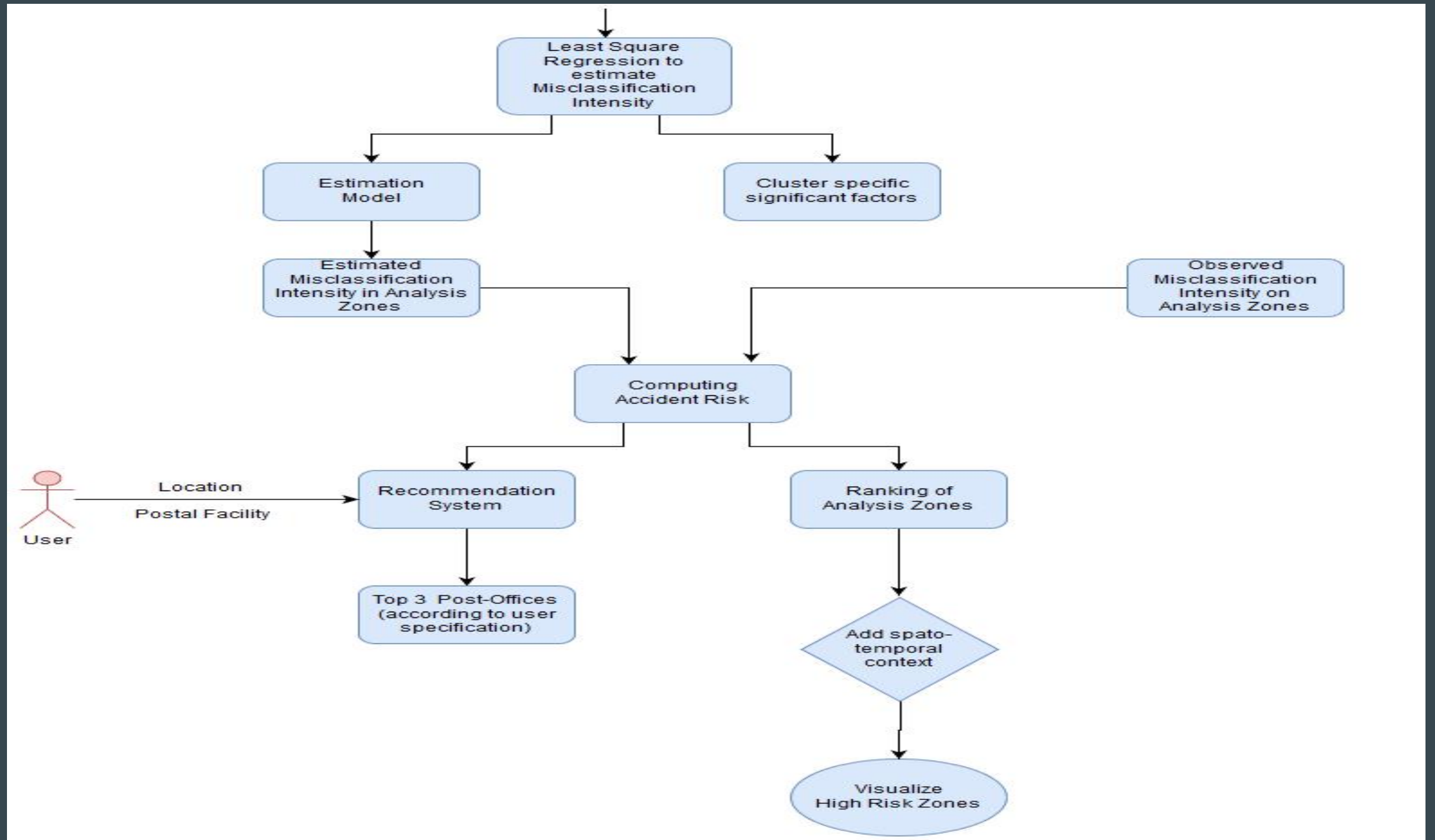
Algorithm :

1. List all the post-offices within a threshold radius from L, which provides F.
2. Rank them in the increasing order of Accident risk and select the top 3.
3. Attach contextual attributes and output the result.

Output : Top 3 post-offices along with their address.

Work Flow Model

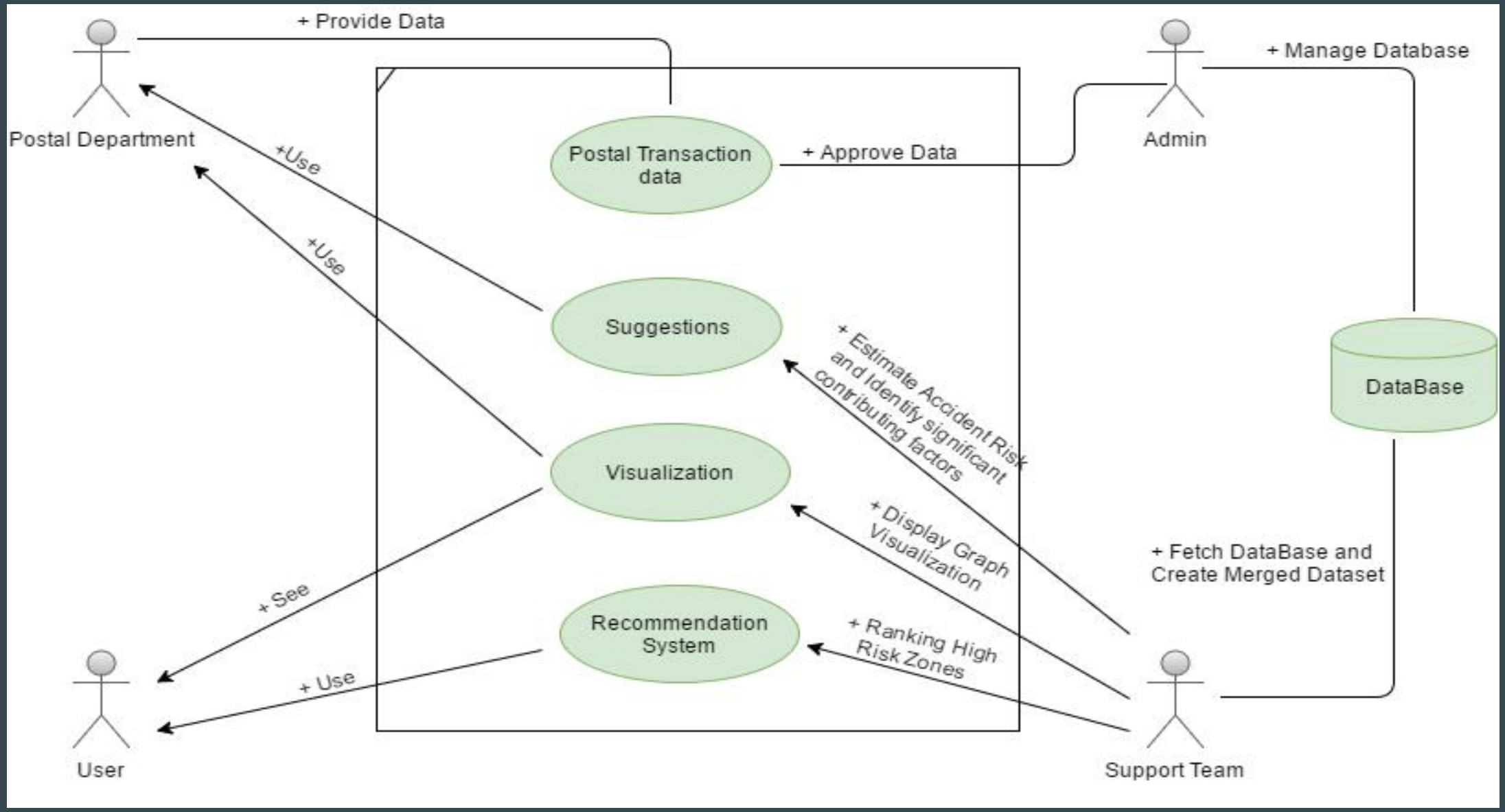



Least Square Regression to estimate Misclassification Intensity
Estimation Model
Cluster specific significant factors
Estimated Misclassification Intensity in Analysis Zones
Observed Misclassification Intensity on Analysis Zones
Computing Accident Risk
Location
Postal Facility
User
Recommendation System
Ranking of Analysis Zones
Top 3 Post-Offices (according to user specification)
Add spato-temporal context
Visualize High Risk Zones

Technology Stack

Apache Spark 2.1.0
MongoDB 3.4.1
Neo4j 3.1.1
R 3.2.2
Python 3.5

Use-Case Diagram



Dependencies

Apache HTTP Server