CAPSTONE PROJECT REPORT

GROUP 10

Recommender System for Best Insurance Provider in Kenya

**Authors:**

* Brian Maina
* Philip Kinga
* Janet Mualuko
* Myles Kiganane
* Purity Murugi

**Date of Submission:** August 2024

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

# **TABLE OF CONTENTS**

### 

[1. Business Understanding](#_g9a2s4onkjjv)

[1.1 Project Objective](#_3v2ptt9gswok)

[1.2 Project Goal](#_2xk20e7zquoq)

[1.3 Target Audience and Motivation](#_31qzcf4doyol)

[2. Data Understanding](#_gft7befprmxu)

[2.1 Data Collection](#_zfdo5nju5ygm)

[2.2 Data Description](#_fefhc9za05qv)

[2.3 Data Sources](#_tksjb1wox9fu)

[3. Data Preparation](#_vza0cdmqczaq)

[3.1 Data Integration](#_fk6cms2z60mj)

[3.2 Data Loading](#_6h5mtj7yg8cu)

[3.3 Data Cleaning](#_5cua0trrlx2d)

[3.4 Challenges](#_rr02oqgimsyp)

[3.5 Saving our Data](#_vxkbcdkstgfi)

[4. Exploratory Data Analysis](#_30l488xf68pl)

[4.1 Statistical Overview of the data](#_8g9b3aaobplr)

[4.2 Univariate Analysis](#_n6ub2xl3eric)

[Visualizations](#_n84xr44ejfjo)

[4.3 Bivariate Analysis](#_qrfkze1qiplc)

[Scatter Plot with Line of Best Fit for Claims Payment vs.Claims Declined ratios:](#_rv8702etmgph)

[Scatter Plot with Line of Best Fit for Claim Papayment vs. Claims closed as no claims ratios:](#_3zjqgv3whw)

[Correlation Coefficients](#_x2yzyanz9bka)

[Linear Regression for Claims payment ratio vs Claims declined ratio:](#_qmgddsxai6q3)

[Linear Regression for Claim Payment Ratio vs. Claims Closed as no Claims Ratio:](#_gbqqjx1ob9am)

[Interpretation:](#_nxw3b8ulwx1s)

[Additional Insights from Linear Regression:](#_4ev2rru2hlgj)

[Correlation Analysis with Claim Payment Ratio:](#_wfjy3t2s8smu)

[Key Observations](#_xipmao9qf1sd)

5. [Feature Engineering, Data Preprocessing, and Scaling](#_g18tmhusrq9r)

[5.1 Data Transformation](#_pwc5n2edzvvw)

[5.2 Data Preprocessing](#_xcbla79u2zyv)

[5.3 Normalizing and Scaling](#_nr7bt8fyfdsq)

[6. Modeling](#_5ms9mpxeccpo)

[6.1 Model Selection](#_ms17n7cosbm2)

[6.2 Target Variables](#_567tu05b0uxg)

[6.3 Model Building and Evaluation](#_8xbyrt67dgei)

[7. Evaluation](#_t5ge5wgqjqoq)

[5.1 Model Performance](#_ci092ur384m2)

[5.2 Justification](#_u40rp680w2yy)

[8. Deployment](#_koev9aucq4mh)

[9. Conclusion](#_s05olovsljbl)

9[.1 Summary of Findings](#_i9kkcw1c848l)

[9.2 Final Model](#_rd2m3niqief4)

[10. Recommendations](#_51cugl7qt9rv)

# **1. Business Understanding**

## 1.1 Project Objective

The project aims to develop a recommender system to help Kenyan customers identify the best non-liability insurance providers based on claim settlement reliability.

## 1.2 Project Goal

The primary goal of this project is to create a system that recommends the most reliable non liability insurance providers to Kenyan customers. By improving the decision-making process, the project aims to increase insurance uptake and customer satisfaction, leading to better financial security for individuals and families.

## 1.3 Target Audience and Motivation

* **Target Audience:** Kenyan individuals and families seeking reliable and affordable non-liability insurance; Insurance companies looking to enhance their services based on customer feedback.
* **Motivation:** The difficulty in finding reliable insurance providers in Kenya has led to low insurance uptake and dissatisfaction. The recommender system aims to address this challenge by simplifying the selection process and providing personalized recommendations.

# **2. Data Understanding**

## 2.1 Data Collection

Data was collected from the Insurance Regulatory Authority (IRA) covering the period from 2018 to 2024, divided into quarterly statistical reports.

## 2.2 Data Description

The data included the following key features:

Identifiers

* Date
* Insurer

Total claims payable features

* Claims outstanding
* Claims intimated
* Claims revived
* Total claims payable

Actionable Claims Data features

* Claims paid
* Claims declined
* Claims closed as no claims
* Total claims action during the quarter

Outstanding claims

* Claims outstanding at the end

Ratios

* Claims declined ratio
* Claims Closed as no Claims
* Claims payment ratio
* Previous quarter Claims payment ratio

## 2.3 Data Sources

Data was sourced from the IRA's quarterly statistical reports.

# **3. Data Preparation**

## 3.1 Data Integration

The data was merged from multiple Excel files into a single file for easier parsing and analysis.

## 3.2 Data Loading

Having combined our data frames externally, we loaded our data onto the notebook.

We checked the headline to confirm we loaded the correct data frame and confirmed it consists of 783 rows and 15 columns.

## 3.3 Data Cleaning

* Dropping irrelevant features such as 'Claim\_payment\_ratio\_ (%)\_previous'
* Strip leading and trailing spaces from column names
* Handling null values
* Handling missing values in Total\_Claims\_Payable, Total\_Claims\_Action\_during\_the\_Quarter and Claims\_outstanding\_at\_the\_end columns
* Data type consistency - Transform all the data in numeric columns into float and then ensure the data in each column is correct.
* Correcting sums for features like Total\_Claims\_Payable, Total\_Claims\_Action\_during\_the\_Quarter and Claims\_outstanding\_at\_the\_end, ratio data.
* Removing duplicates
* Handling outliers for our ratio data
* Handling insurer names into distinct unique values.
* Changing/separating date values using datetime function.

## 3.4 Challenges

The challenges faced during data preparation included incomplete data, duplicates, outliers, and minimal claims data for certain providers in certain quarters.

## 3.5 Saving our data

We saved our data as a .xlsx file for use in the next steps and for future analysis.

# **4. Exploratory Data Analysis**

## 4.1 Statistical overview of the data

The rows provide descriptive statistics including count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum values for each column in the dataset.

## 4.2 Univariate Analysis

To obtain statistics relevant to the data provided and the context of an insurance company/regulator per quarter, we can focus on specific key performance indicators (KPIs) and metrics that are significant in the insurance industry. These include:

* Claims Declined Ratio
* Claims Closed as No Claims Ratio
* Claim Payment Ratio

### Visualizations

We're going to analyze these KPIs for:

* Frequency distribution.
* Top ten companies concerning the average KPI metrics.
* Changes over time.

Frequency Distribution Results

Claims declined and Claims closed as no claims have similar right long tail distribution patterns with most values being around or near zero percent as likely an industry prerequisite where most claims are preferably not declined or left as no claims as the year ends.

Claims declined however had a mmore normalized distribution curve with most claim ratios closed at 40% and most

## 4.3 Bivariate Analysis

For this dataset, we will perform various bivariate analyses to explore how different insurance metrics are related to each other. Here are some of the variables we'll compare:

**Scatter Plots**: To see the relationship between Claims paid vs Claims Declined ratios.

**Correlation Matrix**: To identify and quantify the strength of relationships between our numerical values.

**Pairwise Comparisons**: Using pair plots to visualize pairwise relationships between variables.

**Heatmap**: To visualize the correlation matrix more effectively.

Let's implement these analyses:

### **Scatter Plot with Line of Best Fit for Claims Payment vs.Claims Declined ratios**:

**Purpose**: This plot shows the relationship between the claim payment and the claims declined ratios, with a line of best fit indicating the overall trend.

**Insight**: The line of best fit helps to see whether there is a positive, negative, or no clear relationship between the claim payment and claims declined ratios.

**Interpretation**: Points on the scatter plot represent individual observations, and the red line represents the trend. If the line slopes upwards, it indicates a positive relationship; if it slopes downwards, it indicates a negative relationship.

### **Scatter Plot with Line of Best Fit for Claim Papayment vs. Claims closed as no claims ratios**:

**Purpose**: This plot shows the relationship between the claim payment and the claims closed as no claims ratios, with a line of best fit indicating the overall trend.

**Insight**: The line of best fit helps to see whether paying more claims is associated with having fewer claims closed at the end of the period.

**Interpretation**: Points on the scatter plot represent individual observations, and the red line represents the trend. The slope of the line indicates the nature of the relationship between the two variables.

### Correlation Coefficients

**Claims payment ratio vs Claims declined ratio**: 0.2294840237782586

**Claim Payment Ratio vs. Claims Closed as no Claims Ratio**: -0.1659717921134262

### Linear Regression for Claims payment ratio vs Claims declined ratio:

**Slope**: 0.018820099149975674

**Intercept**: -0.17561527689018275

**R-squared**: 0.05266291716946036

### **Linear Regression for Claim Payment Ratio vs. Claims Closed as no Claims Ratio**:

**Slope**: -0.040902073478427406

**Intercept**: 3.973006333140488

**R-squared**: 0.02754663577734235

#### **Interpretation**:

There is a weak positive correlation between Claims Paid and Claims Declined ratios.

There is a weak negative correlation between Claim Payment and Claims Closed as no Claims ratios.

#### **Additional Insights from Linear Regression**:

For every 1% increase in the Claim Payment Ratio, the Claims Declined Ratio is expected to change by 0.018820099149975674 %

For every 1% increase in the Claim Payment Ratio, the Claims Closed as no Claims Ratio is expected to change by -0.040902073478427406 %

### **Correlation Analysis with Claim Payment Ratio**:

* Moderate positive correlation between Claim Payment Ratio and Claims\_paid (corr = 0.37)
* Moderate positive correlation between Claim Payment Ratio and Total\_Claims\_Action\_during\_the\_Quarter (corr = 0.37)
* Moderate positive correlation between Claim Payment Ratio and Claims\_intimated\_during\_the\_quarter (corr = 0.37)
* Moderate positive correlation between Claim Payment Ratio and Total\_Claims\_Payable (corr = 0.36)
* Weak or no correlation between Claim Payment Ratio and Claims\_outstanding\_at\_the\_beginning\_of\_the\_quarter (corr = 0.28)
* Weak or no correlation between Claim Payment Ratio and Claims\_declined (corr = 0.26)
* Weak or no correlation between Claim Payment Ratio and Claims\_outstanding\_at\_the\_end (corr = 0.26)
* Weak or no correlation between Claim Payment Ratio and Claims\_declined\_ratio\_(%) (corr = 0.23)
* Weak or no correlation between Claim Payment Ratio and Claims\_closed\_as\_no\_claims (corr = -0.03)
* Weak or no correlation between Claim Payment Ratio and Claims\_revived (corr = -0.04)
* Weak or no correlation between Claim Payment Ratio and Claims\_closed\_as\_no\_claims\_ratio (%) (corr = -0.13)
* Weak or no correlation between Claim Payment Ratio and Claims\_closed\_as\_no\_claims\_ratio\_(%) (corr = -0.17)

#### **Key Observations**:

Further Analysis:

* Investigate the features with moderate or strong correlations to understand the underlying relationships.
* Consider potential confounding variables or non-linear relationships that might not be captured by correlation analysis.
* Use these insights to inform feature selection and model development for predicting Claim Payment Ratio.

# **5. Feature Engineering, Data Preprocessing, and Scaling**

## 5.1 Data Transformation

* We're going to perform:
* Label Encoding of the insurer column.
* Quarter Column from our date and year column.
* Creation of our reliablity score.
* Visualize our reliability score.
* Label Encode our reliability score.

## 5.2 Data Preprocessing

This includes splitting the dataset, normalize/standardize the data, performing one hot encoding and label encoding and addressing multicollinearity.

## 5.3 Normalizing and Scaling

Here we scaled our features dataset which was later saved in a separate .csv file.

After scaling our data we have some missing values which we shall handle using the simpleimputer.

# **6. Modeling**

## 6.1 Model Selection

On modeling we focused on developing a robust recommendation system by selecting appropriate modeling techniques and evaluating their performance. Given the complexity of predicting the reliability of non-liability insurance providers, we employed a hybrid approach that combines collaborative filtering and content-based filtering. This strategy leverages both historical claims data and the inherent characteristics of insurance providers to make personalized recommendations

## 6.2 Target Variables

**T**he below target variables were chosen.

* Reliability score (predicting future reliability of an insurance provider)

## 6.3 Model Building and Evaluation

The results of our model was as follows;

### Linear Regression:

* + Root Mean Squared Error: 70.39
  + R-squared: -10.17
  + Mean Absolute Error: 19.60

### Random Forest:

* + Root Mean Squared Error: 7.76
  + R-squared: 0.86
  + Mean Absolute Error: 4.04

### Gradient Boosting:

* + Root Mean Squared Error: 7.20
  + R-squared: 0.88
  + Mean Absolute Error: 3.60

### XGBoost:

* + Root Mean Squared Error: 7.22
  + R-squared: 0.88
  + Mean Absolute Error: 3.54

### Neural Network model

* + Root Mean Squared Error: 2105.81
  + R-squared: -3.75
  + Mean Absolute Error: 13.15

# **7. Evaluation**

## **7.1 Model Performance**

The XGBoost model emerged as the best model with the following metrics:

* Root Mean Squared Error: 7.22
* R-squared: 0.88
* Mean Absolute Error: 3.54

## 7.2 Justification

XGBoost demonstrated strong predictive power, generalization ability, and robustness, making it the ideal choice for helping Kenyan customers identify the most dependable insurance options.

# **8. Deployment**

Our insurance recomender system was deployed both locally and online. The web application was built on Streamlit and its functionality involves selecting the insurer of choice, the year and this displays the reliability score of the insurance company.

<https://kenyan-insurance-provider-recommender-system-wj4p469znhjvmrdvg.streamlit.app/>

### 

# **9. Conclusion**

## 9.1 Summary of Findings

The project successfully developed a recommender system that can improve customer decision-making regarding non-liability insurance, leading to enhanced financial security and customer satisfaction.

## 9.2 Final Model

The XGBoost Regressor was chosen as the best model for predicting the Claim Payment Ratio, achieving the project's goal of developing a reliable recommender system for insurance providers in Kenya.

# **10. Recommendations**

* **Industry Collaboration:** Partner with the Association of Kenya Insurers (AKI) and the Insurance Regulatory Authority (IRA) to obtain more comprehensive data and enhance the model.
* **Consumer Awareness:** Develop educational materials and campaigns to increase consumer awareness and understanding of the rating tool.
* **Further Data Acquisition:** Seek additional data on premium payments, loss ratios, and market share to improve model accuracy.