**CAPSTONE PROJECT REPORT**

**Recommender System for Best Insurance Provider in Kenya Report**

Brian Maina

Philip Kinga

Janet Mualuko

Myles Kiganane

Purity Murugi

**1. Introduction**

The insurance sector in Kenya faces significant challenges, particularly in the non-liability insurance segment, which includes health, property, and personal accident insurance. Many customers struggle to find reliable insurance providers, resulting in low insurance uptake and dissatisfaction. This project aims to develop a recommender system to help Kenyan customers identify the best insurance providers based on claim settlement reliability.

**2. Business Understanding**

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.

**3. Background**

Non-liability insurance is essential for protecting against unforeseen events. However, customers in Kenya often find it difficult to choose a trustworthy insurance provider. This project addresses this issue by simplifying the selection process and providing personalized recommendations.

**4. Domain Coverage**

This project focuses on the non-liability insurance industry in Kenya, specifically targeting the following insurance types:

* Health Insurance
* Property Insurance
* Personal Accident Insurance

**5. Target Audience and Motivation**

The target audience includes:

* Kenyan individuals and families seeking reliable and affordable non-liability insurance.
* Insurance companies looking to enhance their services based on customer feedback.

The motivation is driven by the prevalent difficulty in finding reliable insurance providers in Kenya. The recommender system aims to improve the insurance experience for customers and increase market share for insurance companies by enhancing customer satisfaction.

**6. Expected Outcomes**

The project aims to deliver the following outcomes:

* Enable customers to make informed decisions regarding non-liability insurance.
* Enhance financial security and peace of mind for customers.
* Increase customer satisfaction and market share for insurance companies.

**7. Proposed Methodology**

**Literature Review**

We reviewed existing research on recommender systems, collaborative filtering, and content-based filtering. Additionally, we examined studies on insurance uptake, claim settlement issues, and customer satisfaction in the Kenyan non-liability insurance market. This background research will guide the development of a tailored recommender system for Kenyan insurance customers.

**Data Understanding**

Data will be collected on:

* Claims outstanding
* Claims revived and intimated
* Claims revised
* Claims paid
* Claims declined
* Claims closed as no claims
* Total claims payable
* Total claims action during the quarter
* Claims outstanding at the end
* Claims declined ratio
* Claim payment ratio

**Sources of data include:**

* Insurance Regulatory Authority (IRA) data from 2018 to 2024 divided into quarterly statistical reports.

**The features to be used include:**

* Claims statistics (e.g., outstanding, revived, revised, paid, declined, closed)
* Claims declined ratio, claims closed as no claims ratio, and claim payment ratio for the current and previous quarters.

**Data Preparation**

Conversion of data to excel files

The data obtained from the IRA website was in the form of multiple excel documents for each quarter. This had to be converted into one single excel file to enable easier parsing and upload.

Data will be stored in .xlsx file and will include:

* Insurance provider attributes

**Data preparation steps will involve:**

* Data cleaning (handling missing values, removing duplicates)

One of the features such as Claim\_payment\_ratio\_(%)\_previous, had to be dropped since especially from the first quarter, it would refer to data that wasn’t in the current data.

Total Claims Payable, Total Claims Action during the Quarter and Claims outstanding at the end have many empty entries but could not drop them since they are a sum op other columns, we filled them up using the correct sum.

Finally, most of the features were converted into a float. The date string was converted into a Pandas Datetime.

Duplicates and Outliers could not be removed due to the nature of our data which is into quarters of the same features.

* Normalization (scaling numerical features)
* Encoding categorical variables

**Challenges we experienced include:**

* Incomplete or inconsistent data: Some quarters had missing data due to the structure of our data in quarterly reports.
* Data duplicates: We faced challenges in removing duplicates, which could lead to over-representation of certain records.
* Outliers: Identifying and handling outliers was complex, and all outliers were retained to maintain data integrity.
* Minimal claims data for certain providers: For Trident Insurance and Xplico Insurance, the claims were almost zero to none due to their small market share, which could affect the overall analysis.

Data visualization tools such as histograms, bar charts, scatter plots, heat maps, and correlation matrices will be used for exploratory data analysis.

**Exploratory Data Visualization and Analysis**

In this section, the various insights produced through descriptive statistics and data visualization is presented.

**Data Preprocessing**

Normalization (scaling numerical features) and Encoding categorical variables was applied for our Machine Learning algorithms to work well

**8. Modeling**

A hybrid recommendation approach combining collaborative filtering and content-based filtering will be used. The target variables for the model include:

**For predicting claims frequency:**

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

The baseline model and the rest of the models will use a content based filtering algorithm. The project involves regression (for predicting claim numbers).

**Modeling Results:**

* **Linear Regression:**
  + Mean Squared Error: 4954.49
  + R-squared: -10.17
  + Mean Absolute Error: 19.60
* **Random Forest:**
  + Mean Squared Error: 60.28
  + R-squared: 0.86
  + Mean Absolute Error: 4.04
* **Random Forest with Best Parameters:**
  + Mean Squared Error: 64.25
  + R-squared: 0.85
  + Mean Absolute Error: 4.62
* **Gradient Boosting:**
  + Mean Squared Error: 64.00
  + R-squared: 0.85
  + Mean Absolute Error: 4.53
* **Gradient Boosting with Best Parameters:**
  + Mean Squared Error: 51.83
  + R-squared: 0.88
  + Mean Absolute Error: 3.60
* **XGBoost Model:**
  + Mean Squared Error: 53.91
  + R-squared: 0.87
  + Mean Absolute Error: 3.87
  + Best cross-validation score: 0.95
  + Best Mean Squared Error: 52.17
  + Best R-squared: 0.88
  + Best Mean Absolute Error: 3.54

**9. Conclusion**

This project aims to address the significant challenge of finding reliable non-liability insurance providers in Kenya. By developing a personalized recommender system, we aim to improve the insurance experience for customers and increase market share for insurance companies. The expected outcomes include better decision-making for customers, enhanced financial security, and improved customer satisfaction.

The XGBoost Regressor emerged as the best model for predicting the 'Claim Payment Ratio' and achieving the project's goal of developing a reliable recommender system for insurance providers. It demonstrated strong predictive power, generalization ability, and robustness, making it the ideal choice for helping Kenyan customers identify the most dependable insurance options.