RETIREMENT INCOME PREDICTOR

WHITE PAPER REPORT

BY TEAM 23:

- **STELLA MUKUHI**
- **DANIEL IFEDIBA**
- **UBASINACHI ELEONU**
- **JEFF OUMA**
- **OBINNA UZOEGBU**
- **GEORGE KIBE**

OUTLINE

1. INTRODUCTION

Evidence suggests that we live in an aging society and given the number of people reaching retirement age, there is concern that the majority of the population does not save or invest enough to secure an adequate income in retirement. As this dreadful age for retirement gets near, the level of uncertainty amongst the aging society increases, not only because of lack of savings but other factors which have accumulated over the years.

How can this be managed seeing that most of these factors will remain constant and others may arise?

2. BACKGROUND

Let us think of Jake, a data scientist, who earns over $100,000 annually and has no problem saving and investing a significant portion of his income. Jake has an underlying illness and also channels some chunk of his salary towards medical bills. This sparks some sort of curiosity in Jake as he becomes increasingly interested in knowing the amount of money he should earn every month to enable him to meet his retirement income objective regardless of his illness and what investments would be a contributor to achieving this, bearing in mind the number of dependents he has.

There are a range of factors that make it difficult for Jake to predict his income in retirement. These include:

- ***Employment income:*** How wages contribute to a plan to save for retirement.
- ***Health:*** Uncertainty about if his health status will be a liability to his pension plan.
- ***Retirement age:*** The effect of retirement age on savings, investments and contributions towards his pension plan.
- ***Future financial commitments:*** Number of dependants, the direct and indirect effects dependants will have on disposable his pre and post-retirement.
- ***Financial environment:*** The effects of macroeconomic factors in the financial sector, instability in the housing market and other asset classes that are complementary to pension.

3. PROJECT OBJECTIVES

The team was tasked with predicting the ideal income (the maximum retirement amount that will last until death with a certain confidence level) individuals like Jake must draw from retirement using their personal, investment and retirement data. Our team solved this high level of uncertainty by looking into the individual information provided in the dataset and then analyzing the relationships between each feature and the target variable.

We also looked at the level of importance each feature has in predicting the amount each person should earn to meet their retirement goals. Various regression models were employed to give us a result with high prediction accuracy and fewer errors. The question thus asked is "How much will Jake earn monthly to reach his retirement goal considering various factors that could positively or negatively contribute to this?"

4. OVERVIEW OF THE DATASET

4.1 GENERAL OVERVIEW

The total dataset to be used was made up of the **Train** and **Validation data** which we did our training and metrics evaluation, and then for the **Test data,** we went ahead to perform the actual prediction. The dataset provided consisted of 40 features comprising both categorical and numerical values.

4.2 EXPLORATORY DATA ANALYSIS

Fig 1.1: Total Sample Observations Across all Three Datasets.

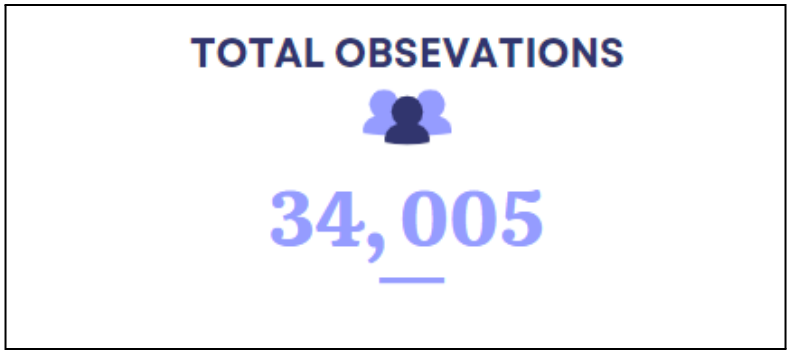


Fig 1.2: The Gender Distribution Across all Three Datasets.

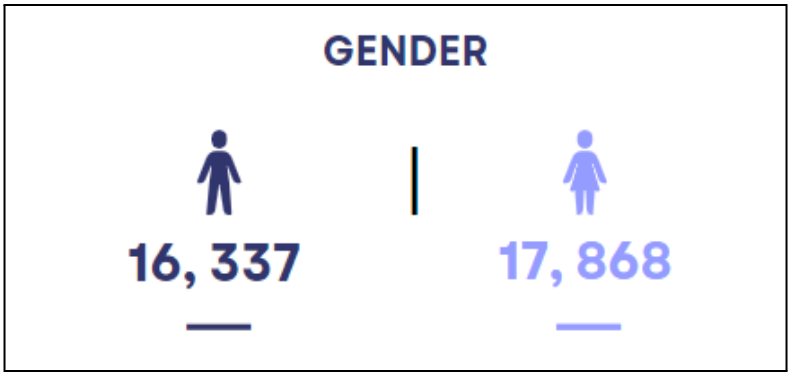


Fig 1.3: Total Amount of People With Emergency Savings.

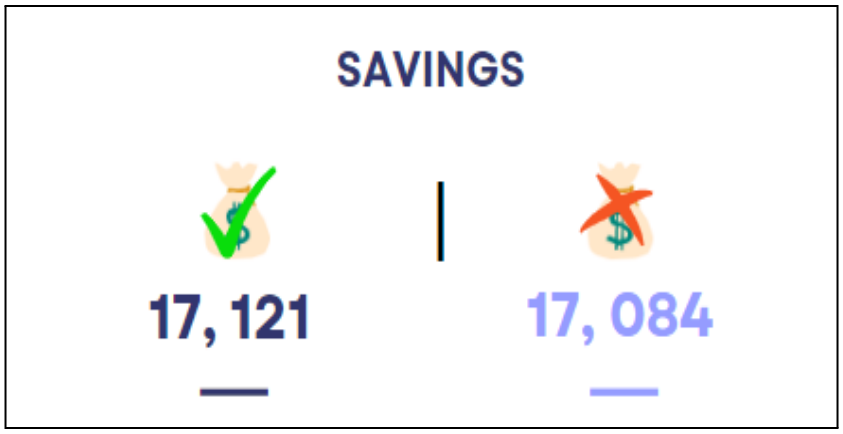


Fig 1.4: The Gender Distribution With Emergency Savings.

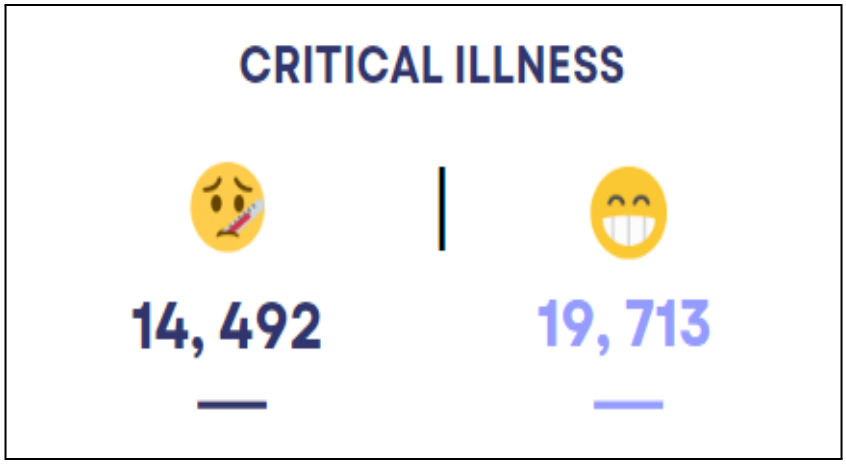
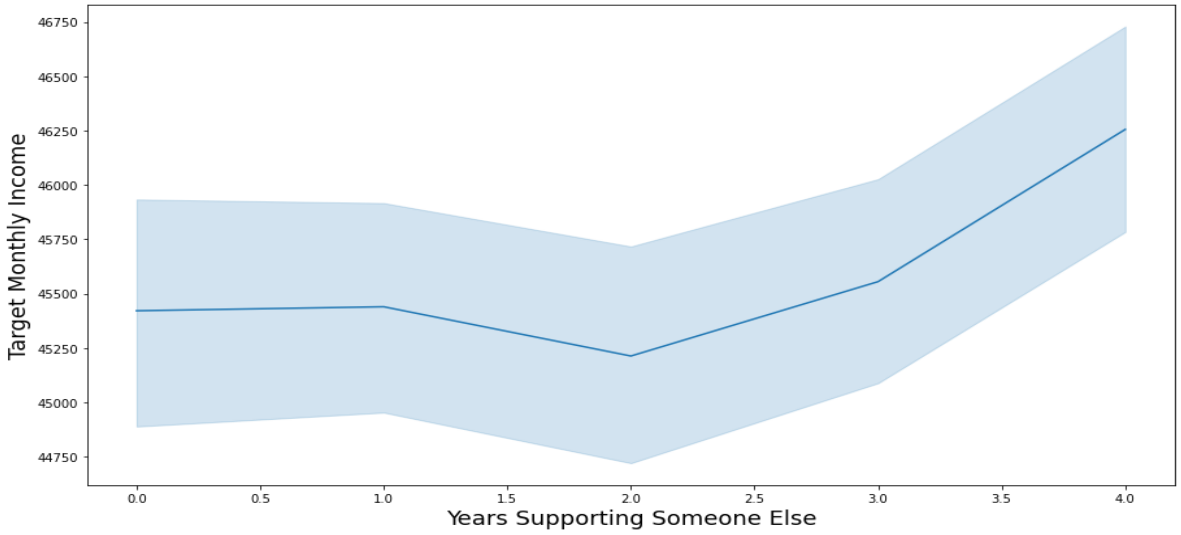


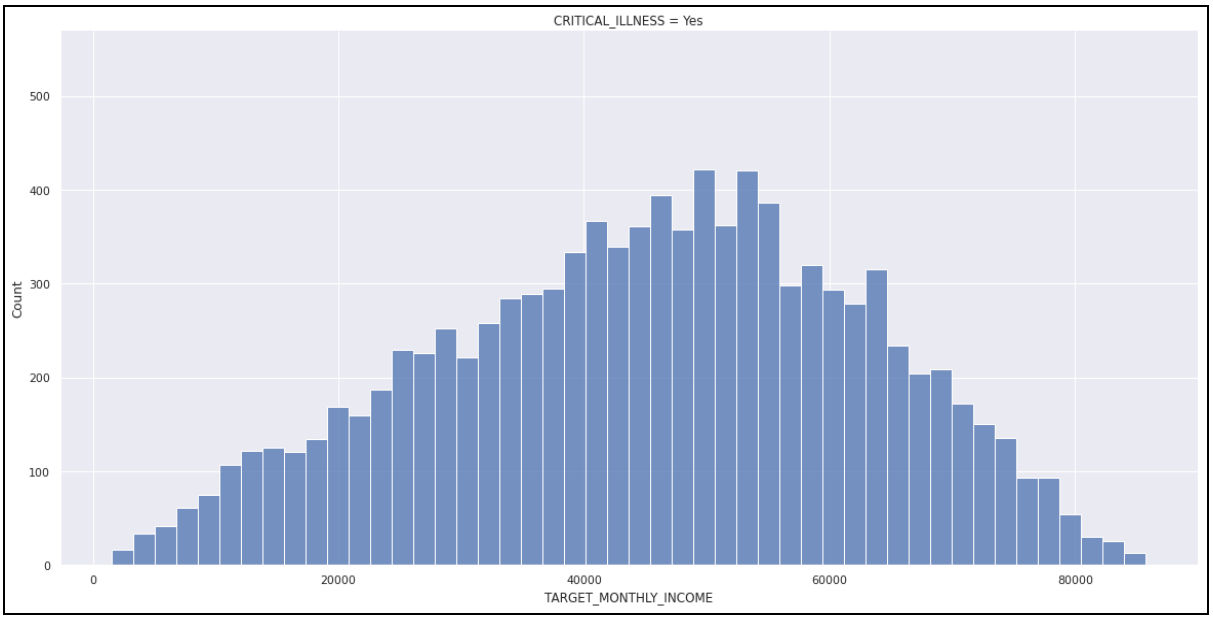
Fig 5: Relationship Between Target Monthly Income & Years Supporting Someone Else



Observations:

This image shows the continous relationship that exists between the years Jake spends supporting someone and the amount he hopes to earn each month to achieve his retirement goal. The same relationship also exists with the years he spends supporting a child.

Fig 6: Relationship Between Target Monthly Income & Critical Illness



Observations:
There seems to be a relationship between how much those with an underlying illness and the amount they are likely to earn due to lack of savings and more medical expenses.

5. APPROACH TAKEN

- ***Label encoding the categorical values****:* Considering the project centers on regression models, it was ideal to convert some categorical values to numerical values by encoding them - changing the feature values to model-compatible format depending on the number of classes in each feature.
- ***Replacing null values with zeros****:* While performing Exploratory Data Analysis on the dataset, and observing that some features in relation to Jake, features such as the spouse retirement age, spouse gender, spouse date of birth contained null (empty) values. In order to prevent less accuracy in the prediction of Jake's income; since it was discovered that there was a null value in his spouse's retirement age and spouse date of birth, this could mean one of two things; he has no partner or there was a computational error while collecting his information. To combat this challenge, we replaced these missing values with zeros to make up for inconsistencies in his information.
- ***Data scaling****:* Taking a deep dive into the statistical analysis of the dataset, using the kurtosis analysis, we deduced that the dataset was prone to abnormalities arising from the presence of outliers. This could result in the features with higher coefficients dominating those with lower coefficients during the prediction phase. To deal with this, the datasets had to be scaled, using the standardization technique, in order to fit the data values between **0 and 1**.

6. MODEL BUILDING & DEPLOYMENT

6.1 MODEL BUILDING

Before building a model that could make predictions for us various factors have to be taken into consideration:

- The error between the predicted values and original values has to be minimized (RMSE – Root Mean Squared Error).
- The model has to neither be Underfitted or Overfitted.
- The model should execute predictions within applicable time frames.

To answer these questions, various machine learning models had to be put to test. Models such as **CATBOOST**, **XGBOOST** and **LIGHT GBM** were implemented to determine the most effective and efficient given the set benchmark performance metric.

Before getting Jake to his desired destination of confidently being able to predict his future, the team encountered some challenges, which were:

6.1.1 OVERFITTING

Regardless of the model in use, the model seems to overfit the data, this had to be tackled so as not to give us the illusion of making wonderful predictions. The following parameters were tweaked to solve the overfitting issue:

- ***Num_leaves****:* LightGBM, unlike XGBoost, uses leaves rather than depth of trees. To reduce overfitting, the lower the num_leaves value, the better. This parameter was reduced to avoid overfitting.
- ***Regularization****:* Be it linear regression, LightGBM, CatBoost or XGBoost, **l1** and **l2** are important parameters for reducing overfitting, they have no specified or recommended value, but the higher the value, the higher the regularization.
- ***Max_bin****:* As the name implies, this refers to the number of bins where we want to fit our predictions. For example; if we have predicted values between ***10 & 80*** and we set the ***max_bin*** value to ***10***, this means that values between ***1 & 8*** go in one bin, ***9 & 16*** go in another and so on. Increasing this can give you accuracy but at the risk of overfitting. The lower the set value for ***max_bin***, the better.
- ***Od_Iter:*** This is a parameter used by the CatBoost model to detect when a model starts overfitting and then stops the training process when overfitting is detected, we use ***early_stopping_rounds*** to tell the model how many iteration it should make before checking for overfitting.

6.1.2 ACCURACY

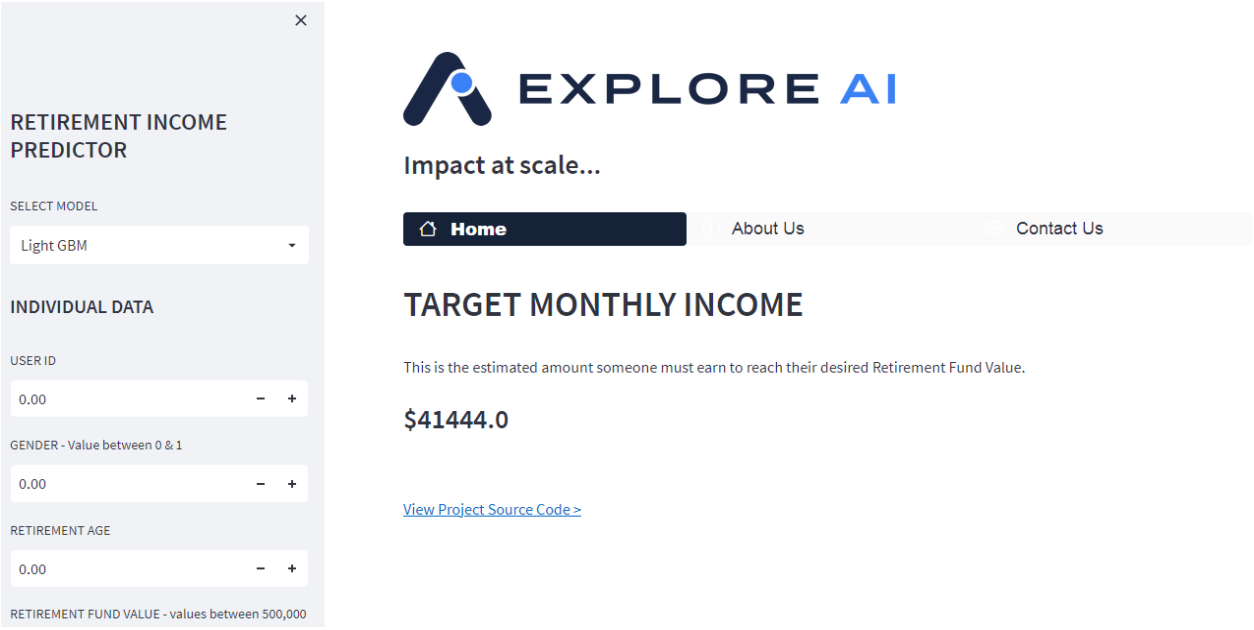
The team struggled with maintaining a high level of accuracy while making sure the model wasn't overfitted. Some parameters using in attaining a high accuracy were:

- ***Learning_rate:*** a tuning parameter in an optimization algorithm that determines the step size at each iteration while moving toward a minimum of a loss function. There is no advised value for this but research has proven that the lower it is, the more accurate the model can be.
- ***Booster:*** The type of model used tells us the type of booster to be used. For LightGBM, ***gbdt*** was used, for XGBoost, ***gbtree,*** and there was no booster used for the CatBoost model.

6.2 MODEL DEPLOYMENT

To make our machine learning process useful to Jake, our team had to deploy these models into a web application with the use of ***streamlit.*** This web deployment service offers a very good and interactive user interface, which our team was able to use.

The web app takes into consideration various models used, the factors that are to be considered in determining one's income and most importantly, how Jake interacts with this app. The app can be visualized below:



METRICS REPORT

MODEL	R^2 SCORE	RMSE	MAE
LIGHT GBM	0.96	3081	2346
XGBOOST	0.95	3198	2531
CATBOOST	0.98	2078	1524

7. CONCLUSION

A result of our weeks-long campaign brought us to the realization that certain factors or features have varying effects on the desired retirement goal of Jake. We were also able to discover investment options that attract optimal yield and are a very viable option for individuals seeking to complement his retirement fund. These includes:

- SA Bond Lap
- SA Cash Lap
- SA Equity Lap
- International Equity Lap

Other non investment options that helps Jake reduce or eliminate this uncertainty are;

- Savings
- Giving extreme focus on his health, as this will result in less medical expenses in the unforeseeable future.

These suggestions don't downplay other investment schemes, but they could yield better returns, as they have more importance to the prediction of one's income.

8. REFERENCES

- Perspectives on Retirement Income Readiness In The United States
- See, e.g., Nevin E. Adams, Rescuing Retirement from the 'Rescuers', NAPA NET (Sept. 27, 2016)
- Stanford - Retirement Income Analysis
- Google Colab Notebook