AI-Powered Lifecycle Financial Planning – Predictive Modeling in Mortality

August 2023

Abstract

How long will you live? This question has extensive implications in the billions of risk estimations made by individuals planning for the future every day. Although never certain, a stronger approximation of an individual's lifespan can enable more reliable future planning and a greater sense of stability than none at all. We reviewed publicly available datasets containing socioeconomic information about U.S. citizens to create a naive model that predicts the likelihood of a person's death at different ages given characteristics such as location, income, place of birth, and more. The results are explained and visualized in this report. While more work must be done to achieve a more accurate predictor, this work provides a baseline for lifespan prediction in coordination with other financial models to aid financial planning.

1 Background

1.1 Dataset Description

For raw data and cleaned data, we use "11 new.csv: NLMS dataset" and "Data.csv: cleaned data with selected features".

This research delves into the age-old question of human lifespan, exploring its implications in countless risk estimations made by individuals planning for their future. While certainty remains elusive, a more accurate approximation of one's lifespan can offer enhanced future planning and a greater sense of stability. In this pursuit, publicly available datasets containing socioeconomic information about U.S. citizens were examined to construct a rudimentary model. This model aims to predict the likelihood of an individual's death at various ages, leveraging characteristics such as location, income, birthplace, and more. The results are dissected and visualized within this report, providing a foundational framework for lifespan prediction, in conjunction with other financial models, to bolster financial planning efforts.

Some of the high-level features we might consider include:

- General biographical information (e.g. age, sex)
- Location
- Occupational and residential environment conditions
- Income
- Medical information

1.2 Dataset Selection

The National Longitudinal Mortality Study (NLMS) [1], curated by the United States Census Bureau, emerged as the optimal dataset for our needs. This dataset was acquired through an approved application process. NLMS offered over 40 features, including key socioeconomic attributes. However, the medical data mainly revolved around causes of death, posing limitations in explaining lifespan prediction. Sample features used in the analysis include age, household size, inflation-adjusted income, income relative to the poverty line, state of residence, urban/rural classification, race, sex, and occupation categories. The dataset employed corresponding weights to account for over/undersampling of specific demographics due to the study's composite nature.

The NLMS dataset contained over 40 features including many of the socioeconomic features that we were looking for. Unfortunately, the medical data provided by the dataset was mostly limited to medical causes of death, which are of dubious explanatory integrity when attempting to predict lifespan. A sample of the features used in the analysis is detailed below:

- Age
- Number of household members
- Inflation-adjusted income
- Income relative to poverty line
- State of residence and urban/rural classification
- Race
- Sex
- Occupation categories

[1] https://www.census.gov/topics/research/nlms.html

1.3 Preprocessing

The dataset undergoes various stages of cleaning and transformation to ensure it is suitable for further model building. Firstly, a preliminary examination of its shape and potential missing values is performed. The dataset is narrowed down to focus on individual-level data, specifically targeting instances where there is only one borrower.

The exploration of missing values in various features, such as borrower's race, gender, age, credit score, and ethnicity, highlights particular variables that might require imputation. A meticulous elimination of missing values is executed to sustain data integrity. Furthermore, redundancy is addressed by removing variables that exhibit no variation, thereby ensuring that the dataset is devoid of multi-collinearity and irrelevant features for predictive modeling.

Subsequent steps involve the removal of potentially redundant. We need to ensure that the data is not only clean but also optimal in terms of the features included in further analyses. Specific transformations, such as converting categorical variables into a numerical format, are applied to facilitate the application of machine learning algorithms in subsequent steps.

The data is inspected to ensure that the variables are of the correct type (e.g., categorical), preserving the clean and processed data, which can ensure that our predictive modeling are based on optimized data.

2 EDA

2.1 Target Variable

The NLMS data is a classification dataset that 1787583 observations of people health condition, more specifically, alive or not. The target variable ”indmort” is a binary value, 0 stands for a man is alive and 1 stands for dead.

Here is a count of the binary target variable.

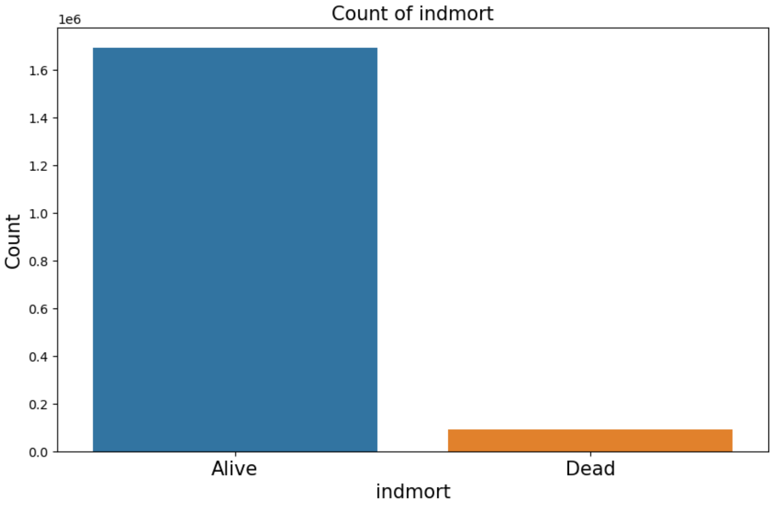


Figure 1: Count of indmort

From the plot we could see that 0 is a huge part in the observations, more specifically, the ratio between 0 and 1 is 18 : 1. The skewed distribution implies that the dataset may suffer from class imbalance, which might be challenging for us to predict the minority class (dead class) accurately.

2.2 Feature Variable

The clean data we used for the model has **32** features out of the target variable.

In order to facilitate the analysis, we list these variables by their meanings.

Variable Name	Description
age	Age at Time of Interview
hhnum	The number of persons residing in the household at the time of the interview
povpct	Income as Percent of Poverty Level
adjinc	Inflation Adjusted Income
stater	State Recode
pob	Region of Birth
sex	Sex, 1 = male, 2 = female
race	Race, 1 = White, 2 = Black, 3 = American Indian or Eskimo, 4 = Asian or Pacific Islander, 5 = Other nonwhite
urban	Urban/Rural Status, 1 = Urban, 2 = Rural
smsast	SMSA Status based on county boundaries, 1 = SMSA, in central city; 2 = SMSA, not in central city; 3 = Not in an SMSA
wt	Adjusted Weight
indmort	Death Indicator, 1 = Dead, 0 = Alive

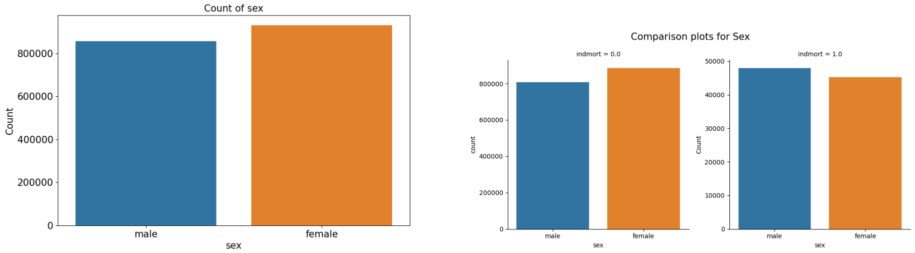
In order to have a clear look about the relationship between target variable and feature variables, we pick several examples from each group and split them between categorical and numerical variables.

2.2.1 Categorical Variable

Sex

We get the distribution of "sex" variable and the count plot of "sex" across different classes of "indmort". For good comparison, we cat two plots with two target classes together. In this data, we have more females than males.

From the comparison plot, we could find that the "sex" class change significantly with the target classes. Among people alive, there are more females than males, while among people dead, there are more males than females. This shows that "sex" might be a good predictor of the target.



(a) Count of sex

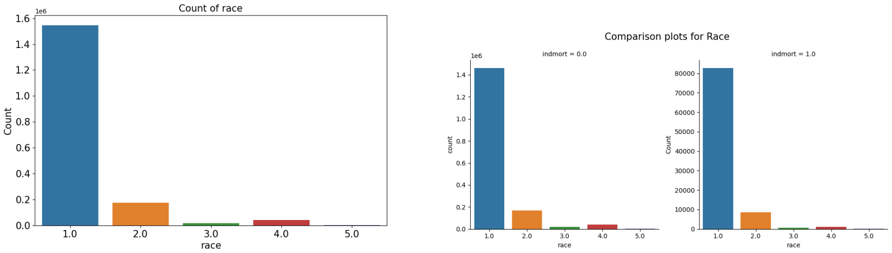
(b) Comparison plots for Sex

Figure 2: Sex plots

Race

We get the distribution of “race” variable and the count plot of “race” across different classes of “indmort”. For good comparison, we cat two plots with two target classes together. In this data, most observations’ races are in group 1.0, which is White.

From the comparison plot, we could find that the “race” class do not change much depending on the target, although a little bit difference in the height of Group 3.0 and 4.0. This shows that race might not be a huge influence in people’s mortality.



(a) Count of race

(b) Comparison plots for Race

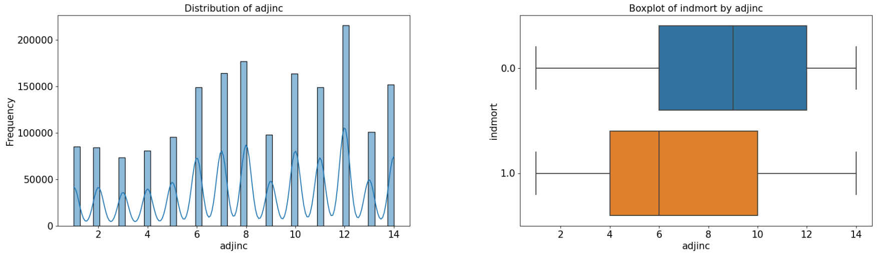
Figure 3: Race plots

2.2.2 Numerical Variable

Adjinc

We get the distribution of “adjinc” variable and the boxplot of “adjinc” across different classes of “indmort”. In this data, “adjinc” is distributed normally like but with a strange decease in 9 and an increase in 12.

From the comparison boxplot, we could see a strange difference in the distribution across different target classes. The median of people alive (indmort = 0.0) is higher than the median of people dead (indmort = 1.0), implies that, on average, people with high income tend to be alive. This plot reminds us that “adjinc” might be a good predictor for “indmort”.



(a) Distribution of adjinc

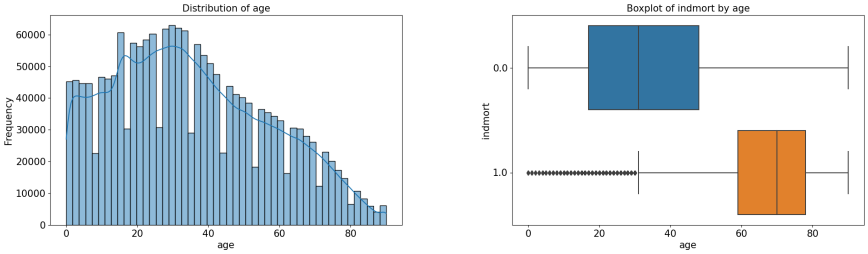
(b) Boxplot of indmort by adjinc

Figure 4: Adjinc plots

Age

We get the distribution of “age” variable and the boxplot of “age” across different classes of “indmort”. In this data, “age” is distributed right-skewed. This suggests that there is a higher concentration of data in the center of the distribution with a few extreme values on the right. This means that there are relatively few people with very high age, while the majority of individuals have lower to moderate ages.

From the comparison boxplot, we could see a strange difference in the distribution across different target classes. The median of people dead (indmort = 1.0) is higher than the median of people alive (indmort = 0.0), implies that, on average, people with higher age tend to be dead. The height of box for people dead is narrower than boxes for people alive, showing that the ages for people dead are more tightly clustered around the median. The whiskers for people dead is shorter than those for people alive, it indicates that dead people’s age have fewer extreme expenditures. Age for people dead has more outliers than age for people alive, it suggests that age for people dead may have more people with exceptionally low ages. This plot reminds us that “age” might be a good predictor for “indmort”.



(a) Distribution of age

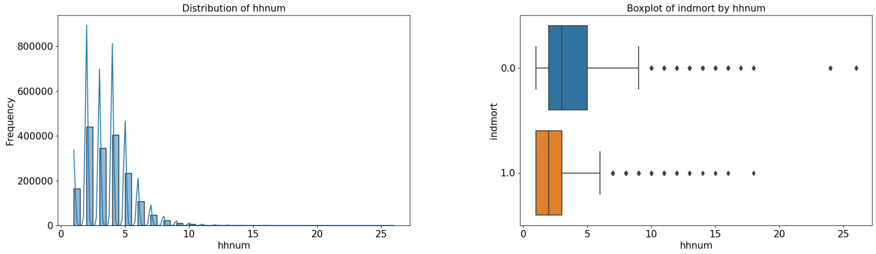
(b) Boxplot of indmort by age

Figure 5: Age plots

Hhnum

We get the distribution of “hhnum” variable and the boxplot of “hhnum” across different classes of “indmort”. In this data, “hhnum” is distributed right-skewed. This suggests that there is a higher concentration of data in the center of the distribution with a few extreme values on the right. This means that there are relatively few families with very high amount of people, while the majority of families have lower to moderate amount of people.

From the comparison boxplot, we could see a difference in the distribution across different target classes. The median of people alive (indmort = 0.0) is higher than the median of people dead (indmort = 1.0), implies that, on average, people in a family with more members tend to be alive. The height of box for people dead is narrower than boxes for people alive, showing that the amount of people in the family of people dead are more tightly clustered around the median. The whiskers for people dead is shorter than those for people alive, it indicates that the amount of dead people’s family members have fewer extreme expenditures. The amount of family members for people alive has more outliers than those for people dead, it suggests that the amount of family members for people alive may have exceptionally high numbers. This plot reminds us that “hhnum” might be a good predictor for “indmort”.



(a) Distribution of hhnum

(b) Boxplot of indmort by hhnum

Figure 6: Hhnum plots

2.3 Heat Map

We also create a heat map to check the correlation between each numerical variables.

Note that the correaltion between numerical variables are not high, but we find some signficant examples that make normal sense.

“adjinc” and “povpct” has a very high correlation of 0.90. It is because these two variables are computed together as “adjinc” is the amount of income while “povpct” is the income as percent of poverty level.

“hhnum” and “age” has a high negative correlation of -0.46. It implies that older people a tend of has fewer family members, which might not the same in normal society has we belive people with higher ages should have more children and the amount of family members should be high.

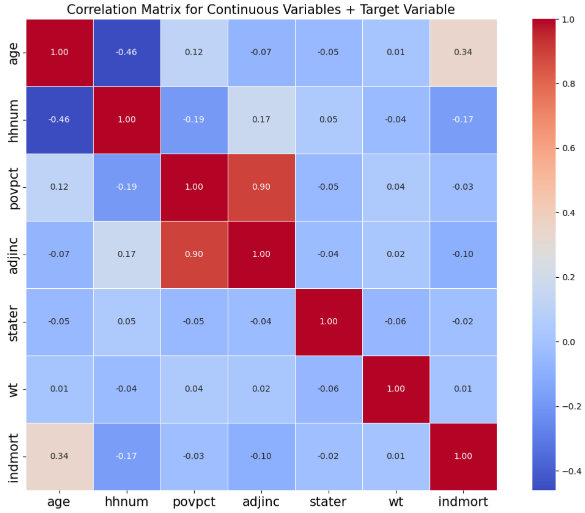


Figure 7: Heat map for Numerical Variables

3 Modeling

3.1 Oversampling

The mortality data from our dataset was naturally highly imbalanced. Over the 10-year follow-up period, there existed a ratio of approximately 19 survivals for each death. This can cause binary classification models to underfit the data and simply classify every survival and death entry as a survival, therefore reaching a 0.95 accuracy. To combat this, we used the Synthetic Minority Oversampling Technique (SMOTE) from the **imblearn** library. After splitting the data into a large training and small testing subset, the technique is applied only to the training data. This resamples and imputes the training data so that it shifts to a 1:1 death to survival ratio, allowing the classification model to more accurately discern deaths when scored on the test data.

3.2 Models

3.2.1 Logistic Regression

Logistic regression is a type of statistical model often used for classification and predictive analytic. Logistic regression estimates the probability of an event occurring, which is dead and alive in our dataset. More specifically, we are using binary logistic regression as our target variable has only 2 possible outcomes, 0 and 1. We use *LogisticRegressionCV* from the *scikitlearn* package.

3.2.2 LightGBM

The LightGBM classifier is a tree-based model that outputs binary probabilities. The logistic regression is a linear based standard model with parameters specified in the code. We use $LGBMClassifier$ from the $LightGBM$ package. The model was trained using cross validation and hyperparameter optimization over a feature space. The models were trained over 150 iterations of the optimization.

3.3 Evaluation Metrics

When evaluating the performance of a classification model, there are several metrics we can use to assess how well the model is performing. The choice of metric(s) depends on the specific characteristics of the problem and what aspects of the model's performance are most important. For this data set we use these evaluation metrics:

1. **Accuracy**: Accuracy is the most straightforward metric and represents the ratio of correctly predicted instances to the total number of instances in the dataset. It's a good metric for balanced datasets but can be misleading when dealing with imbalanced datasets.

   Accuracy = (Number of Correct Predictions) / (Total Number of Predictions)

2. **Precision**: Precision measures the ratio of true positive predictions to the total number of positive predictions made by the model. It's a good metric when the cost of false positives is high.

   Precision = (True Positives) / (True Positives + False Positives)

3. **Sensitivity**: A key performance metric in classification problems. It measures the ability of a classification model to correctly identify all positive instances in a dataset.

   Sensitivity = (True Positives) / (True Positives + False Negatives)

4. **Specificity**: important performance metric in classification problems, especially when we want to assess a model's ability to correctly identify negative instances.

   Specificity = (True Negatives) / (True Negatives + False Positives)

5. **Confusion Matrix**: A confusion matrix provides a tabular representation of the model's predictions, showing the number of true positives, true negatives, false positives, and false negatives. It's a useful tool for gaining a detailed understanding of a model's performance.

3.4 Results

Here is the result for each model:

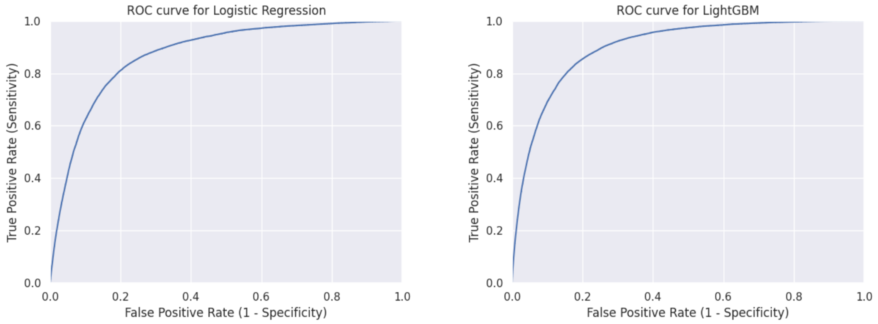
Model	Logistic Regression	LightGBM
Accuracy	0.8277	0.9369
Precision	0.1997	0.3976
Sensitivity	0.7696	0.4159
Specificity	0.8308	0.9654

Table 1: Result

Note that LightGBM has a better accuracy and precision than Logistic regression.

3.5 ROC Curve

The ROC (Receiver Operating Characteristic) curve is a graphical representation of the model's performance across different thresholds. It plots the true positive rate (TPR or recall) against the false positive rate (FPR) at various threshold settings.



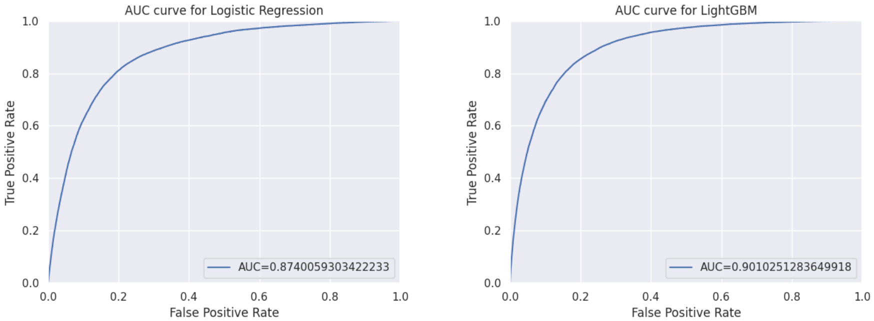
(a) ROC Curve for Logistic Regression

(b) ROC Curve for LightGBM

Figure 8: ROC Curve

3.6 AUC Curve

AUC-ROC (Area Under the ROC Curve) quantifies the overall performance of the model across different threshold settings. It provides a single value that represents the model's ability to distinguish between classes. A higher AUC indicates better performance.



(a) AUC Curve for Logistic Regression

(b) AUC Curve for LightGBM

Figure 9: AUC Curve