Reddit Survey on Financial Independence

Overview:

The dataset is from Reddit Survey and represents the financial independence of individuals and other variables that related to it such as person's curent assets, liablities, income etc. and the data does not include retired individuals.

The purpose of the model is to understand the factors that contribute to whether a person would feel financially independent or not?
The dataset has 1998 rows and 65 variables.

The following link gives a clear description of all the outcome variables :

https://www.openintro.org/data/index.php?data=reddit_finance

The purpose of this project is to model for Financial Independence of a person using certain Input Variables.

Data Cleaning:

Data Cleaning had to be done for this project. Any ASSUMPTIONS in the cleaning process has been put in Block Letters.

1. Feature Engineering:

    - All expenses related columns were combined to make a separate column total_exp. Since expenses remain similar for a person over the years, it is sensible to combine to a more consistent variable. Variables combined were :

    (‘2020 housing expenditure, 2020 utilities expenditure, 2020 transport expenditure‘, 2020 necessities expenditure, 2020 luxury expenditure, 2020 child expenditure, 2020 debt repayment, 2020 charity, 2020 healthcare expenditure, 2020 taxes paid, 2020 educational expenses, 2020 other expenditure‘)

    - All Debts/Liablities related columns were combined to make a separate column total_liablities. This is crucial because a person would only feel financially independent considering all his liablities together and hence giving the model a more consistent variable.

      (‘student loans‘, ‘mortgage‘, ‘auto loan‘, ‘credit personal loan‘,‘medical debt‘, ‘Debt from Investment Properties‘, ‘Other Debt’)
    - All Assets/Holdings related columns were combined to make a separate column total_asssets. This is crucial because a person would only feel financially independent considering all his assets and net worth together. Variables combined were :

      (‘home value‘, ‘brokerage accounts tax‘, ‘retirement accts tax‘, ‘cash‘, ‘invsestment accounts‘, ‘crypto investment‘, ‘Investment in Properties / businesses‘, ‘Other Assets‘)
    - The four columns representing full time status, part time status, gig status and not employed status”represents the work status of an individual. The data collection was poor for this as these could have been combined into one column. Hence, these variables were combined to create a separate column ’`work_status'`. So, combining would make one consistent column with different work statuses. The final column would limit values to three different types : full time, part time, personal gig, unemployed.
    - For the variable representing ‘relationship status’, it had 6 distinct values which could be combined to either ‘Single’ or ‘in a Relationship‘.

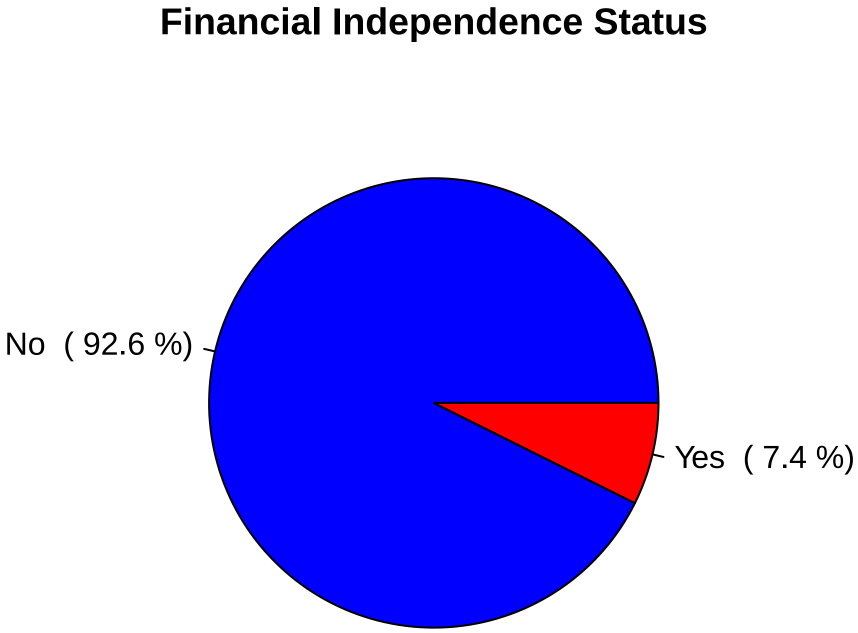
- The column ‘children’ had 4 entries - 1. ‘Do not have children but intend to’ 2. ‘Have children’ (Both combined to one entry ‘Yes’) 3. ‘N/A’ and 4. ‘Do not have children and do not intend to’ (Both combined to one entry ‘No’). This is done **ASSUMING** people who have children or intend to have them will have different financial independence number than people who don’t have children.

2. The age(current age) and retire_age(represents at what age person wants to retire) variable was in the format of a range like ‘21-25’. The mean value was taken of this range to be used to provide a numerical variable to the model and mean would give an average output of each range of ages.

3. For the column ‘Country’, there were 54 different countries and all data is represented in their individual currency, creating inconsistency. USA, UK, Australia and Canada holds 97% of the values in the data and are amongst the developed countries. So all the data of other countries apart from the mentioned ones were removed.
   We **ASSUME** here that USA, UK, Australia, Canada are similarly developed nations.

4. For the purpose of our model, the gender values were combined into 3 distinct values - Male, Female and Others for consistency.

5. **MISSING VALUES:** We removed the all the rows with n/a values: Total expenditure , Total worth , Total liablities , 2020 investment saving, 2020 gross income, retiring age, current age , children, gender and Number of Income Earning Members



Modeling:

1. Our outcome variable - ‘fin_indy’ which represents whether the person is financially independent or not.

2. We used **logistic regression** for our model here. Since our results will be either in the form of Yes or No, i.e. if a person feels financially independent or not, *the outcome is binary* it becomes sense to use logistic regression for our problem.

3. For our first model - As per priori selection, we took the following variables for our first model :
   total expenditure, total worth, total liablities, 2020 investment saving, 2020 gross income, retiring age, current age, retire expenditure, Annual Income from Retired Assets, Target Safe Withdrawl Rate, Investment Amount for Retirement, Percent Financial Independence achieved, Children , Gender Number of Income Earning Members .
   The p-values for all of these came out to be very high, greater than 0.95 and hence we removed a few variables to focus on the more essential variables.

4. For our second model - we used the following variables :
   Total expenditure (numerical), Total worth (numerical), Total liablities (numerical) , 2020 investment saving‘(numerical), 2020 gross income(numerical), retiring age(numerical), current age (numerical) , children (categorical), gender (categorical) and Number of Income Earning Members (numerical)

   REASON TO SELECT THE VARIABLES : These seemed to be highly correlated to a person considering him as financially independent. For instance, a person who has a lot of total_worth and less total Liablities would be more prone to feeling financially independent. Similarly, a person’s current age would would be correlated to how much savings, assets would a person own and being financially independent.

Model 1:

After performing the first model, we got a result where all the P-values were greater than 0.95. The model couldn’t indicate any strong relationship between being Financially Independent and other variables. So we had to reduce the variables to make a better model.

Model 2 & Cook’s Distance :

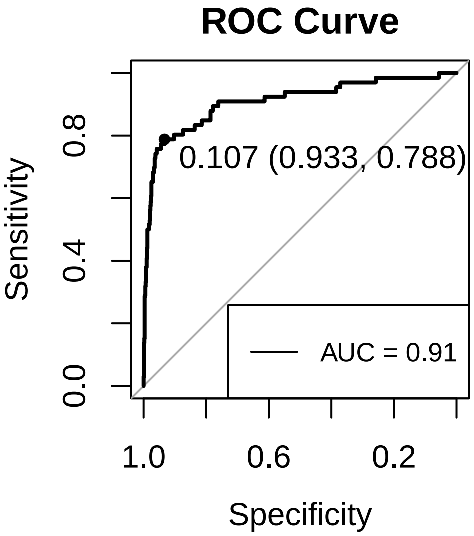
Our model 2 gives us strong relationships between various variables with Financial Independence.

The plot above shows that the one data point was highly influencing our model. We wanted to run another model after removing this influence point and see the results. The point of discussion here with clients is to understand the reason for that data point influencing our model.

Final Model Output :

```
Call:
glm(formula = factor(fin_indy) ~ total_exp + total_worth + total_liablities +
    `2020_invst_save` + `2020_gross_inc` + retire_age + age +
    factor(children) + factor(gender) + num_incomes, family = "binomial",
    data = reddit_finance_supersub)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          1.441e-01  1.052e+00   0.137 0.891066
total_exp           -2.298e-06  3.100e-06  -0.741 0.458626
total_worth          1.653e-06  2.256e-07   7.327 2.36e-13 ***
total_liablities    -2.819e-06  7.990e-07  -3.528 0.000419 ***
`2020_invst_save`   -1.950e-06  2.490e-06  -0.783 0.433439
`2020_gross_inc`    -7.377e-07  2.086e-06  -0.354 0.723653
retire_age          -1.414e-01  3.149e-02  -4.490 7.11e-06 ***
age                  9.271e-02  3.291e-02   2.817 0.004846 **
factor(children)Yes -4.322e-01  3.303e-01  -1.309 0.190658
factor(gender)Male  -6.902e-01  3.980e-01  -1.734 0.082867 .
factor(gender)Other -3.479e+01  1.290e+04  -0.003 0.997847
num_incomes2        -1.170e+00  3.776e-01  -3.098 0.001945 **
num_incomes3         2.155e+00  2.343e+00   0.920 0.357803
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 523.05  on 1309  degrees of freedom
Residual deviance: 310.20  on 1297  degrees of freedom
AIC: 336.2

Number of Fisher Scoring iterations: 8
```



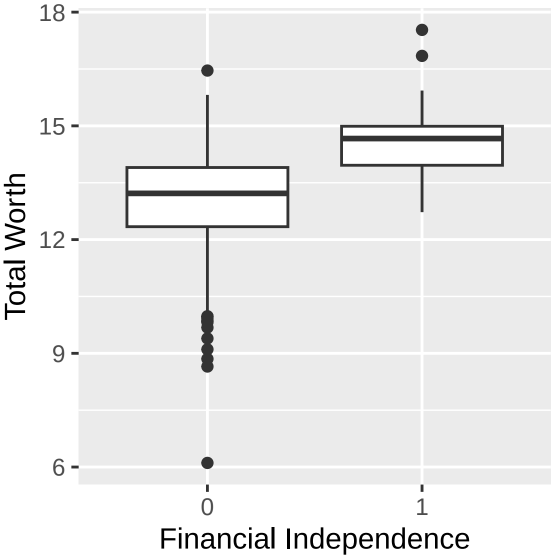
Model Interpretation:

1. Our final model gives us strong results where we found multiple input variables highly correlated with the output variable of Financial Independence. These include - Net Worth, Total Liablities, Retiring Age, Current Age, Gender and Number of Income Earning Members.
2. We have used AUC metric to interpret the results. The model's AUC comes as 0.91, showing strong performance of the classification model.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1160   14
         1   84   52

               Accuracy : 0.9252
                 95% CI : (0.9096, 0.9389)
    No Information Rate : 0.9496
    P-Value [Acc > NIR] : 0.9999

                  Kappa : 0.4795

 Mcnemar's Test P-Value : 3.168e-12

            Sensitivity : 0.78788
            Specificity : 0.93248
         Pos Pred Value : 0.38235
         Neg Pred Value : 0.98807
              Precision : 0.38235
                 Recall : 0.78788
                     F1 : 0.51485
             Prevalence : 0.05038
         Detection Rate : 0.03969
   Detection Prevalence : 0.10382
      Balanced Accuracy : 0.86018

       'Positive' Class : 1
```

Results :

1. The sensitivity of the model is 0.79, indicating that model is able to predict 79% of the true positive instances compared to all positive instances.
2. With each additional total worth, the log odds of Financial Independence increases by 1.653 * 10^-6, all else held constant.
3. With each additional increase in Age, the log odds of Financial Independence increases by 0.092, all else held constant.
4. The sensitivity of the model is 0.93, indicating that model is able to predict 93% of the true negative instances compared to all negative instances.
5. Our Kappa value comes out to be 0.48, suggesting that there is a 48% agreement beyond what would be expected by chance alone. This indicates a moderate level of agreement between the observers or raters involved in the study.



4. The plot above compare Total Worth with Financial Independence. It is clearly observed that there's a clear difference between people's worth between the two classes, validating our model's output that Net Worth is strongly correlated with Financial Independence.

Conclusions :

1. We can conclude that variables Net Worth, Total Liablities, Retiring Age, Current Age, Gender and Number of Income Earning Members are the factors that contribute to whether someone considers themselves as financially independent.

Future Work :

There's an issue with data imbalance. Around 93% of the people represents the data supporting they arent financially independent. Hence, for future more data needs to be collected for a better model.