UNLOCKING HAPPINESS: WHAT DRIVES HAPPINESS IN SOUTH AFRICA?

INTRODUCTION

South Africa is a nation famous for its rich diversity, with over 63 million people belonging to a variety of cultural, linguistic, and socio-economic backgrounds. As a result of this diversity, there are **significant disparities in how different groups perceive their quality of life**. We expect that wealth would be a factor that is highly associated with quality of life. However, it is by no means clear what the **specific commodities** are, that the people of South Africa value most. Understanding the factors that contribute to happiness is essential in capturing the needs of this diverse population. By examining these factors, **we aim to identify what people in various communities value most in their lives.**

Unfortunately, **poverty** in South Africa remains a significant issue, with over **49% of the population living below the upper-bound poverty line**. It ranks low on global happiness indices, with a 2024 ranking of **83rd out of 143 countries** in the World Happiness Report, and 110th out of 193 countries in the Human Development Index. We hope that our findings will play a small role in helping guide the **strategic allocation of resources and funding**, ensuring that investments are targeted where they will have the **greatest impact on improving the quality of life for South Africans.**

OBJECTIVE

To identify factors that are **significantly associated with happiness among South Africans**, so that its resources might be allocated effectively to improve the quality of life of its citizens.

THE DATA SET

The General Household Survey by Stats SA is an annual household survey which measures key areas like health, housing, food security, and agriculture, making it ideal for studying socio-economic dynamics in South Africa.

Our findings focus on the 2022 Survey. Future research could apply the same methodology to other datasets, for a more comprehensive analysis.

Limitations

- Potential **sampling biases** due to underrepresentation in rural or informal settlements,
- Reliance on **self-reported** data
- **Cross-sectional** nature - can't establish causality, only association

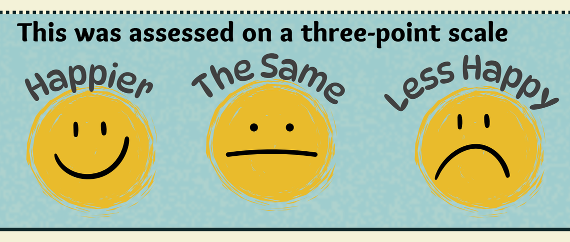
Sampling Procedure

- The survey uses **stratified two-stage sampling** and structured questionnaires, covering all de jure household members, ensuring the sample size is large enough to be representative of the South African population.
- However it excludes collective living quarters such as student hostels, old age homes etc.

Variable of Interest

Measures a **household's perceived happiness** with life relative to how they felt ten years ago.

This was assessed on a three-point scale



Limitations: Variable of Interest

- The dataset lacks a baseline happiness measure, limiting our ability to account for initial differences.
- The subjective nature of the response variable introduces inherent noise, affecting accuracy.

METHODOLOGY

01. Data Cleaning

- **Removed** any missing data or non-respondents to reduce noise in the model.
- Converted the variable of interest from a 3-point scale into a **binary response**:
  1 = improved happiness
  0 = no improvement in happiness
  to enable the construction of a **logistic regression model.**

02. Stepwise Regression

- Employed **stepwise regression** to select variables with a significant impact on the response variable.
- Used forward selection and backward elimination methods to refine the predictor set.
- This approach helped identify key variables, but is **limited by certain assumptions**

03. Random Forests

- **Explored Random Forests:** A flexible machine learning method that handles complex interactions between predictors without assuming linear relationships.
- Assessed variable importance using the **Gini Out-of-Bag (OOB) index.**

✓ **Important Variables:** Higher OOB Gini values → Help improve the model

✗ **Less Useful Variables:** Zero or negative OOB Gini values → May cause the model to be too specific, reducing accuracy.

04. Model Refinement

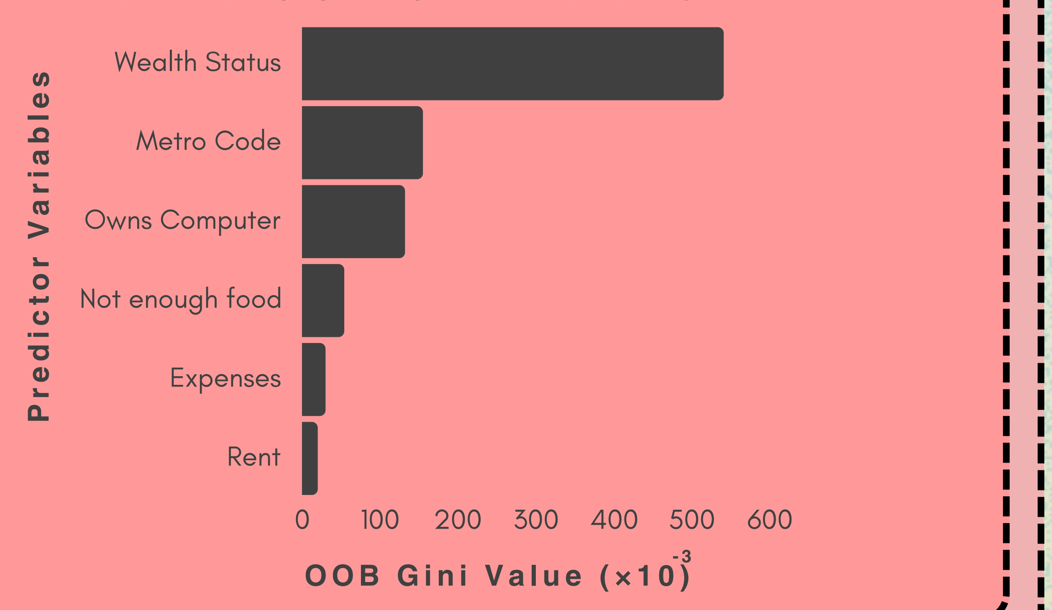
- **Predictors Selected:** Chose those with a positive OOB Gini value for the logistic regression model.
- **Model Comparison:** Compared Random Forest model (C-Index: 0.72) with Stepwise model (C-Index: 0.65)

C-Index values closer to 1 indicate better prediction accuracy

- **Final Choice:** Selected Random Forest-based logistic regression for better predictive performance.

RESULTS

The most significant variables according to the OOB Gini values



After identifying all the predictors with positive OOB Gini values, **predictors were iteratively removed** until a further removal would have caused a significant decrease in predictive ability of the model. This resulted in a **final model using the above 6 predictors.**

Summary Statistics of the Model

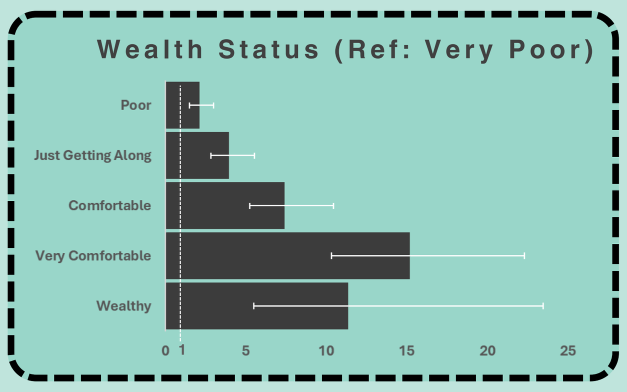
Statistic	Value	Conclusion
AUC-ROC	0.72	The model has moderate predictive ability*
Lemeshow GoF	p = 0.36	The model fits the data well

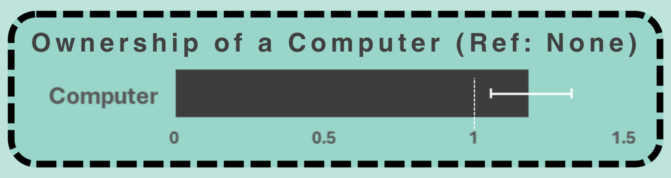
*considering the subjective nature of a variable measuring happiness[1]

Confusion Matrix of the Model

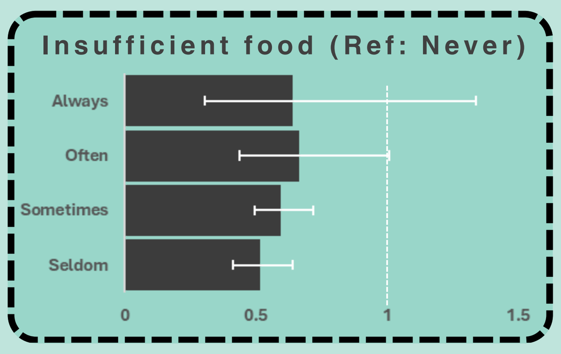
9.78% True +ve	12.68% False +ve
13.25% False -ve	64.29% True -ve

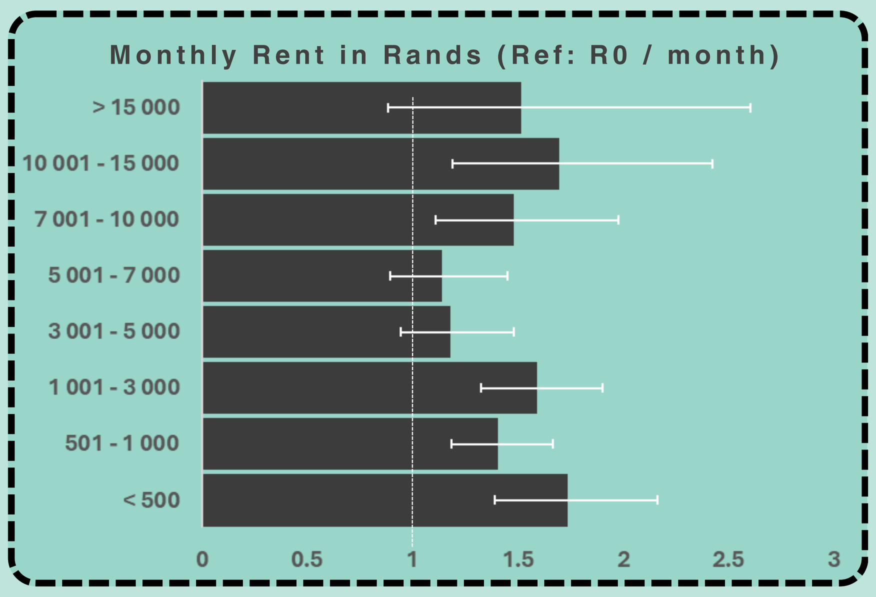
95% Confidence Interval Plots for the odds ratios of the predictor variables





Note: A 95% confidence interval that does not contain 1, implies statistical significance at a 5% level of significance





- Since each predictor is categorical, **the odds ratios are ratios of the odds of an improved happiness rating, for the categories on the y-axis versus the reference category given in the chart heading.**
- The length of the bar represents the point estimate, with the white lines representing the confidence intervals.
- Due to the large number of categories in the Expenses and Metro Code predictor variables, charts for these are omitted. However, we mention the relevant conclusions below.

CONCLUSIONS

Wealth as a Primary Driver

Wealthier households reported higher likelihoods of improved happiness, reinforcing the link between financial security and well-being.

Geographical Impact

Urbanized areas, such as Cape Town and Nelson Mandela Bay, showed higher happiness levels compared to non-metropolitan areas, highlighting regional disparities in infrastructure and opportunities.

Financial Burdens

High expenses and rent appear to be negatively associated with happiness, indicating the toll of financial pressure on well-being.

Food Security Matters

Households facing hunger were less likely to report happiness, underscoring the importance of nutrition in life satisfaction.

Significance of Computer Ownership

Households owning computers were 1.2 times more likely to be happy, potentially due to enhanced access to education, job opportunities and connectivity.

Unexplained predictors

Variables that we hypothesized to have an effect on happiness, such as the source of financial income or access to healthy food, did not show significant results in the model. Although, this could also be attributed to other confounding factors not captured in the model, or the complex nature of subjective well-being.

Caution

Subjective data like happiness is shaped by biases, social desirability, and is susceptible to influence by transient emotional states, rather than long-term life satisfaction, leading to variability in the data. This introduced complexities into achieving a model with high accuracy and concordance.

Research Value

While our findings may seem intuitive, they validate commonly held assumptions with data, provide insight into regional dynamics and identify potentially underexplored predictors like computer ownership. These results can inform evidence-based polices and serve as a foundation for future studies to explore trends overtime.

REFERENCES:

[1] Trescato I. Roversi C. Vettoretti M. Di Camillo B. (2022). A predictive model of subjective wellbeing in the elderly population. Retrieved from https://www.researchgate.net/publication/365391827_A_predictive_model_of_subjective_wellbeing_in_the_elderly_population. **[2]** Delsignore G. Olivan-Blazquez B. Aguilar-Latorre A. (2021). Measuring happiness in the social sciences: An overview. Retrieved from https://www.researchgate.net/publication/349490505_Measuring_happiness_in_the_social_sciences_An_overview **[3]** Nyongesa D. (2020). SAS Global Forum: Variable Selection Using Random Forests in SAS. Retrieved from https://support.sas.com/resources/papers/proceedings20/4826-2020.pdf **[4]** Helliwell J. Huang H. Shiplett H. Wang S. (2024). World Happiness Report 2024. Retrieved from https://worldhappiness.report/ed/2024 **[5]** World Population Review (2024). Human Development Index (HDI) by Country 2024. Retrieved from https://worldpopulationreview.com/country-rankings/hdi-by-country **[6]** Statistics South Africa (2019). Five facts about poverty in South Africa | Statistics South Africa. Retrieved from https://www.statssa.gov.za/?p=12075 **[7]** Statistics South Africa. (2022). General Household Survey (GHS). Retrieved from https://www.statssa.gov.za/?page_id=16405