Analysis and modelling of the depression dataset.

Lawrence R

Problem, Key Findings and Next Steps

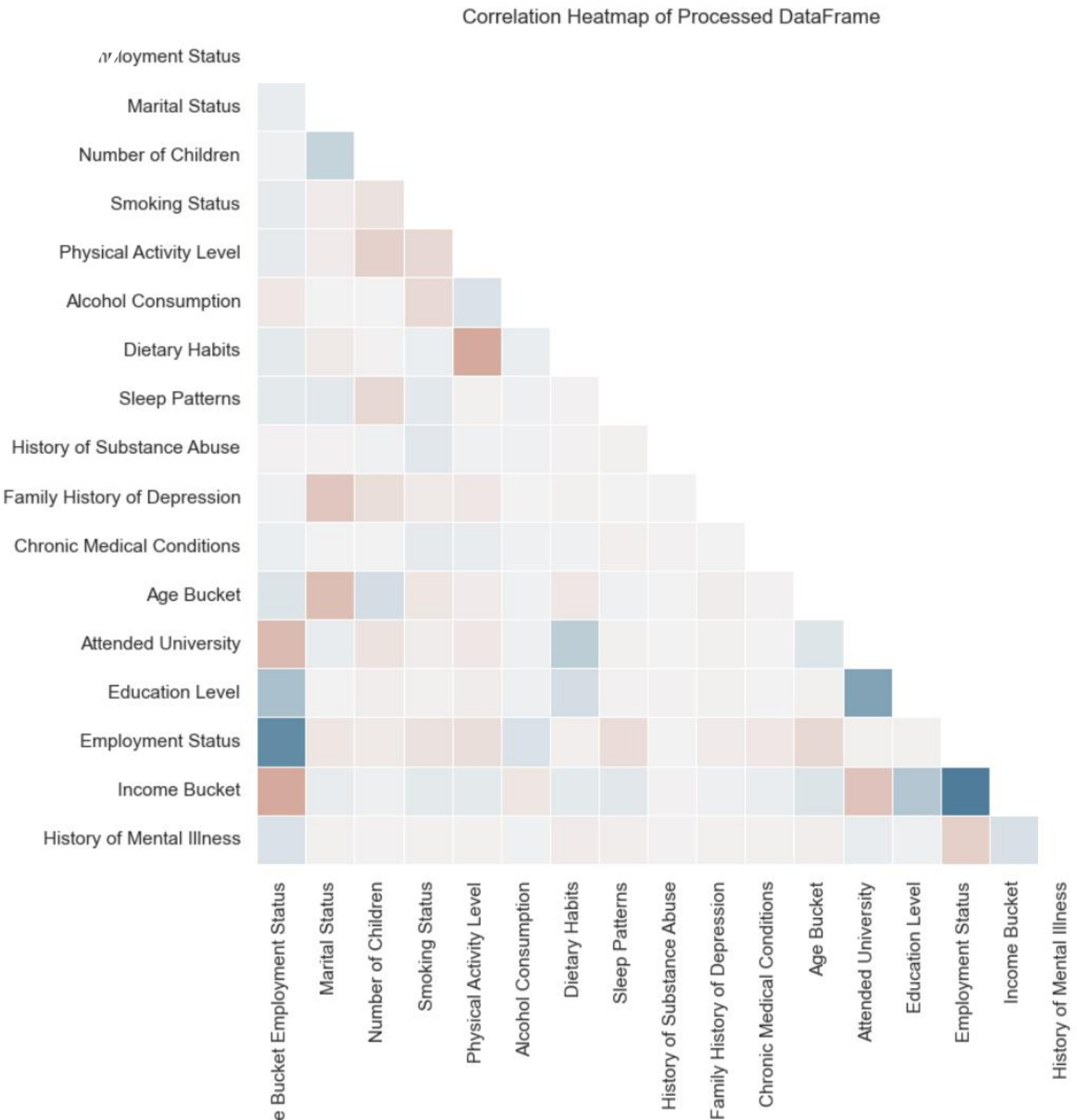
- **Goal:** Predict individuals at risk of depression using lifestyle and socioeconomic factors.
- Focused on **recall** to catch as many at-risk individuals as possible, even if accuracy was lower.
- Dealt with imbalanced target 'History of Mental Illness' using model sample weights.

Model Performance (CatBoost, Weighted for Recall):

- Accuracy: **47.76%** **-** Recall: **80.92%** **-** ROC-AUC: **57.1%** **-** Precision: **34.64%** **-** F1 Score: **48.51%**

Main Insights from the Data:

- **Income, employment status, and education** were the strongest predictors.
- **Employment and income were correlated**, but kept separate for better performance.



Problem, Key Findings and Next Steps

SHAP Analysis - Feature Importance

Most Important Features:

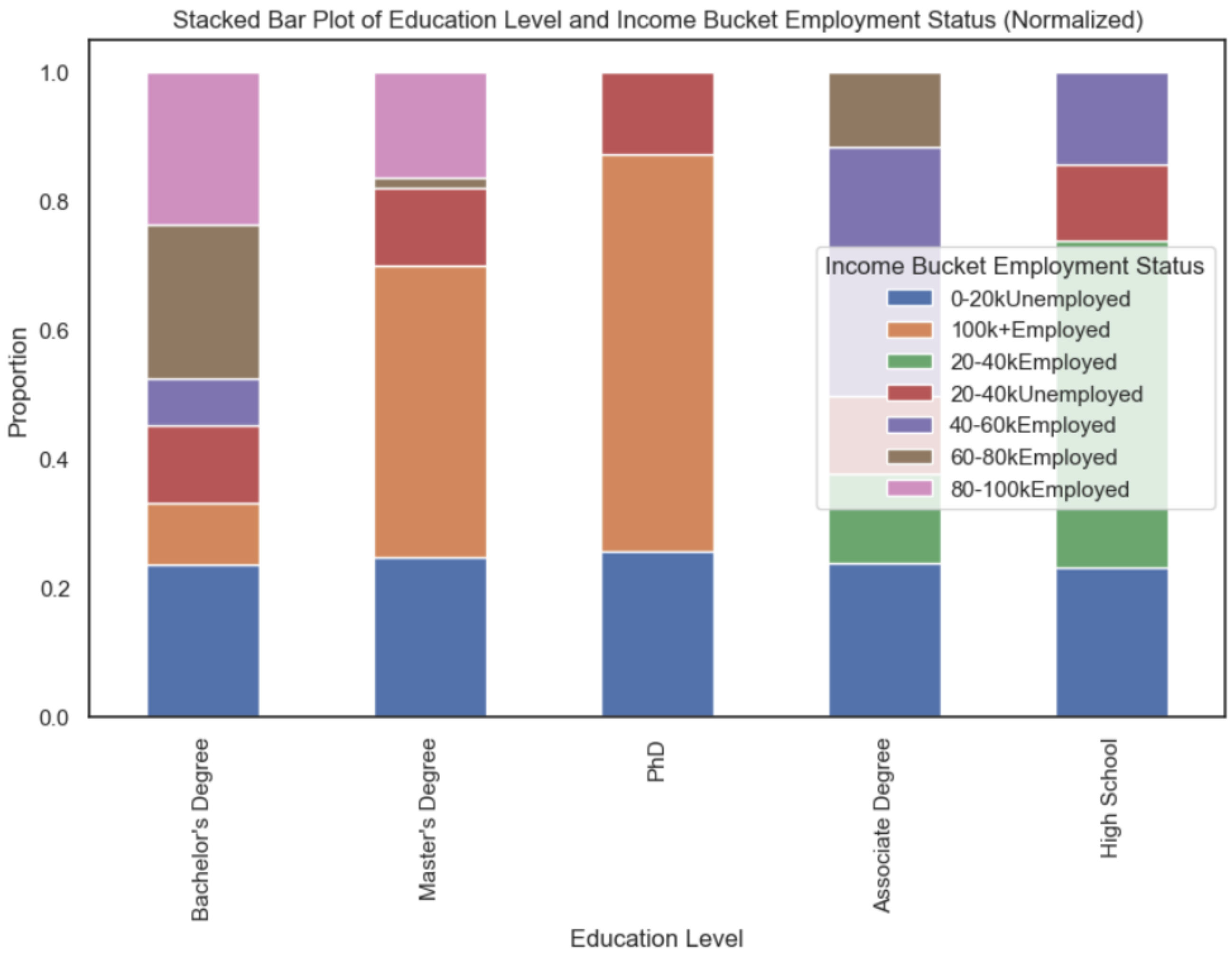
- Income and Employment Status (highest impact)
- Education Level and Smoking Status

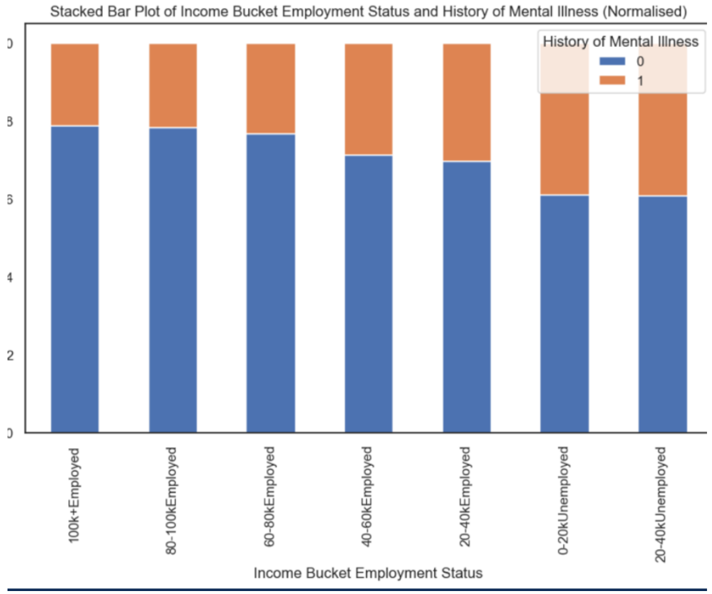
Key SHAP Insights:

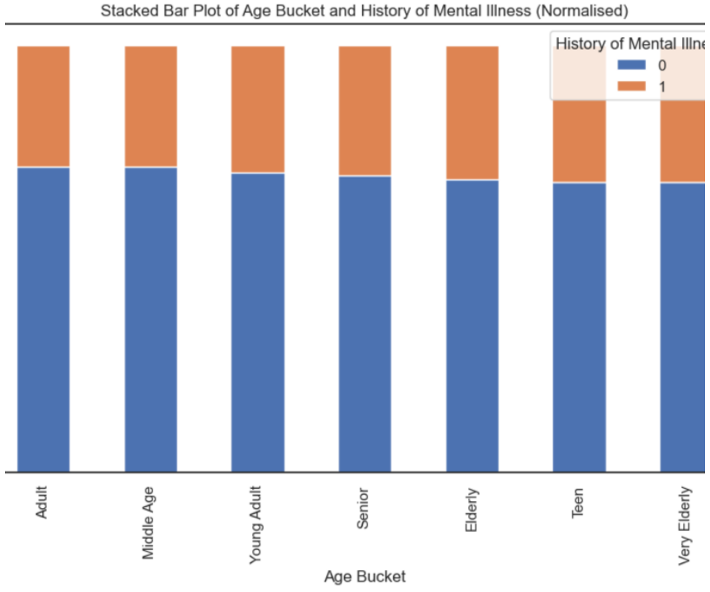
- Lower income and unemployment increase the risk of depression.
- Higher smoking frequency correlates with higher depression risk.

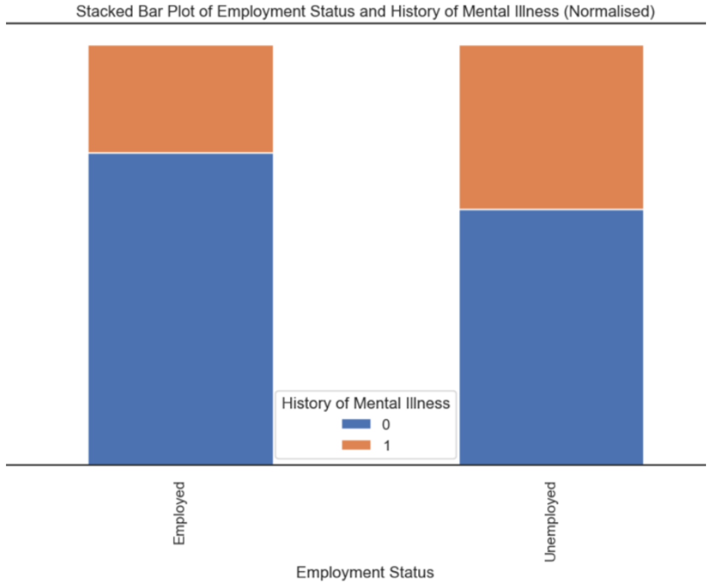
Further Work & Next Steps

- Improve Model Performance
  - Test feature interactions (e.g., combine employment and income).
  - Try ensemble models (stacking CatBoost, LightGBM, AdaBoost).
  - Further tune hyperparameters.
- Enhance Interpretability
  - Use SHAP visualisations to explain predictions.
  - Create an interactive dashboard for stakeholders.
- Address Data Biases & Limitations
  - Investigate class imbalance (resampling, synthetic data).
  - Check for demographic biases in predictions.
- Real-World Deployment Considerations
  - Monitor model drift over time.
  - Test in real-world settings to evaluate impact.









Key Findings from Data Analysis

- **Employment status, income, and education** were **strong predictors**.
- **Employment & income were correlated** (higher income = more stable employment).
- **Sedentary lifestyle & unhealthy diet** were also correlated but left as separate factors.

History of Mental Illness

Categorical

Distinct	2
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	404.3 KiB

History of Mental Illness (Target Variable)

- **Binary classification:** Predicting whether an individual has a history of mental illness (**Yes/No**).
- **Imbalanced dataset:** More **0s (no history)** than **1s (history of mental illness)**, making it harder to detect positive cases.
- **Balancing Techniques:** Could use **undersampling** (removing majority class samples) or **oversampling** (duplicating minority class samples), but…
- **Our Approach:** Used **sample weight scaling** instead, ensuring the model **focused more on positive cases** without artificially modifying the data.
- **Key Features:** Strong correlations with **income, employment status, education level, and smoking habits**.
- **Model Focus: Maximised recall** to **reduce false negatives** and capture as many at-risk individuals as possible.

Model & Approach

Why CatBoost?

- Works well with **categorical data** (most of the dataset).
- **Fast on GPU**, allowing efficient Optuna hyperparameter tuning.
- Handles **imbalanced data** well using the **scale_pos_weight** parameter.
- **Hyperparameter Tuning with Optuna**
- Tuned **iterations, depth, learning rate, and more** to optimise recall.
- **Sample weight tuning** helped push predictions towards identifying **at-risk individuals**.

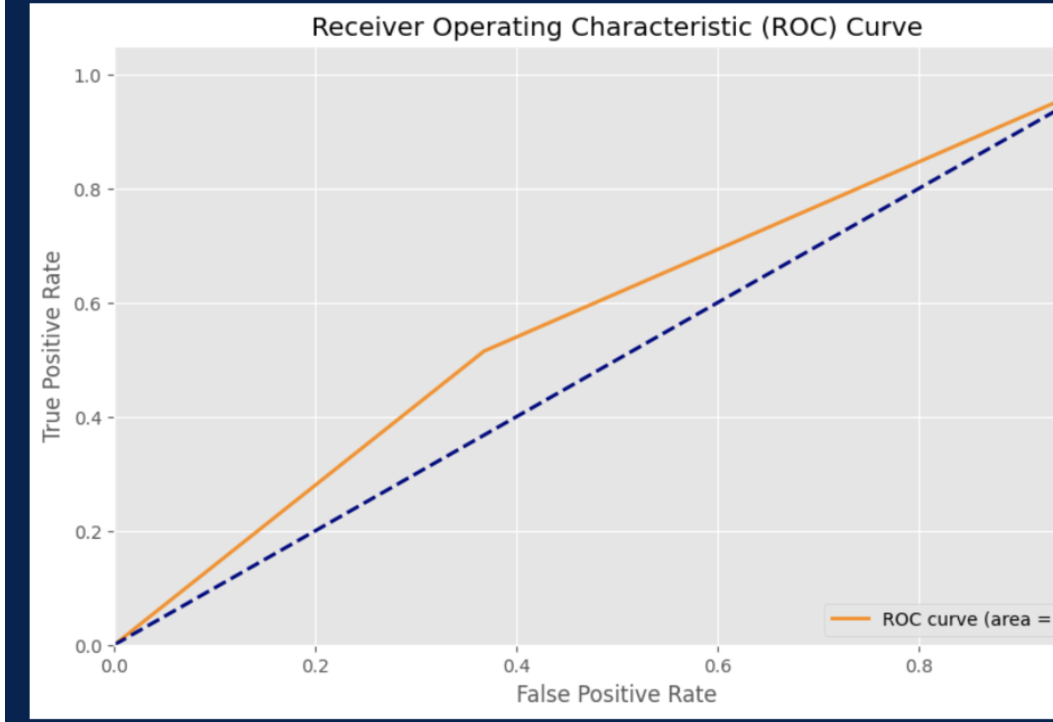
Performance Metrics & Focus on Recall

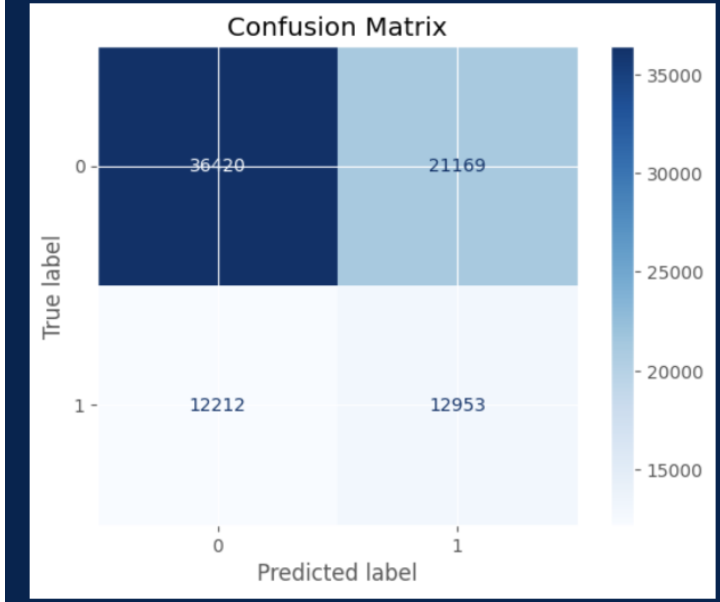
- **Recall was prioritised** to **avoid missing individuals at risk of depression**.
- Trade-offs between **AUC-ROC, F1-score, Precision, and Recall**, but **recall was key**.

Modelling Results- normal weights

- The first model produced fair results but precision and accuracy of the model could be better.
- We would like to over predict depression cases so we don't miss cases of depression.
- We can increase the sample weights to do this.

	Accuracy	Recall	Roc_Auc	Precision	F1 Score
Catboost_normal_weight	0.5966	0.5147	0.5736	0.3796	0.437

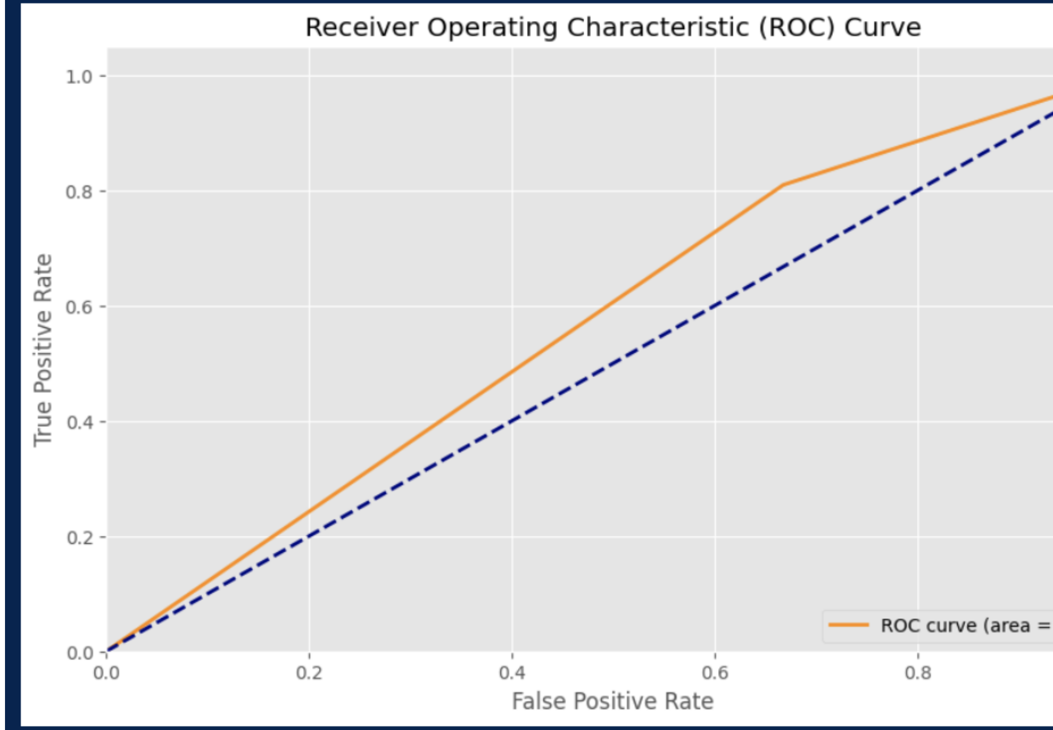


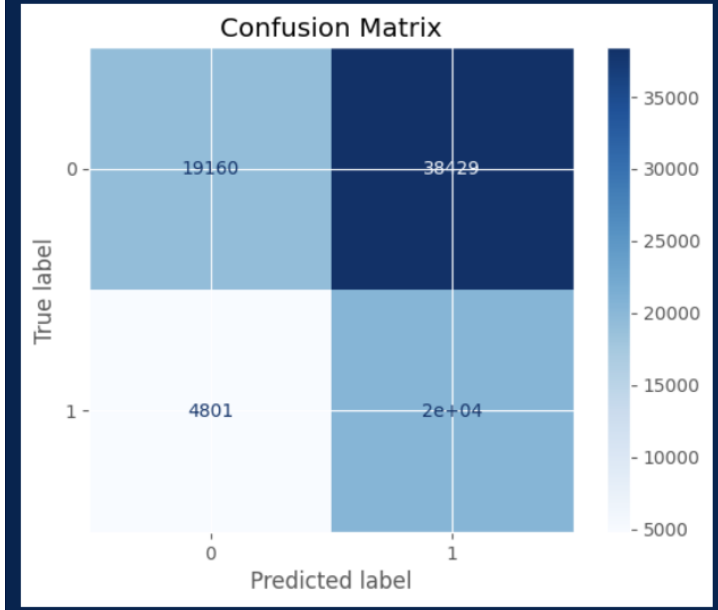


Modelling Results – extra weights

- By adding extra weights we reduce the cases of depression missed but increase the false detections of depression sacrificing our Accuracy.

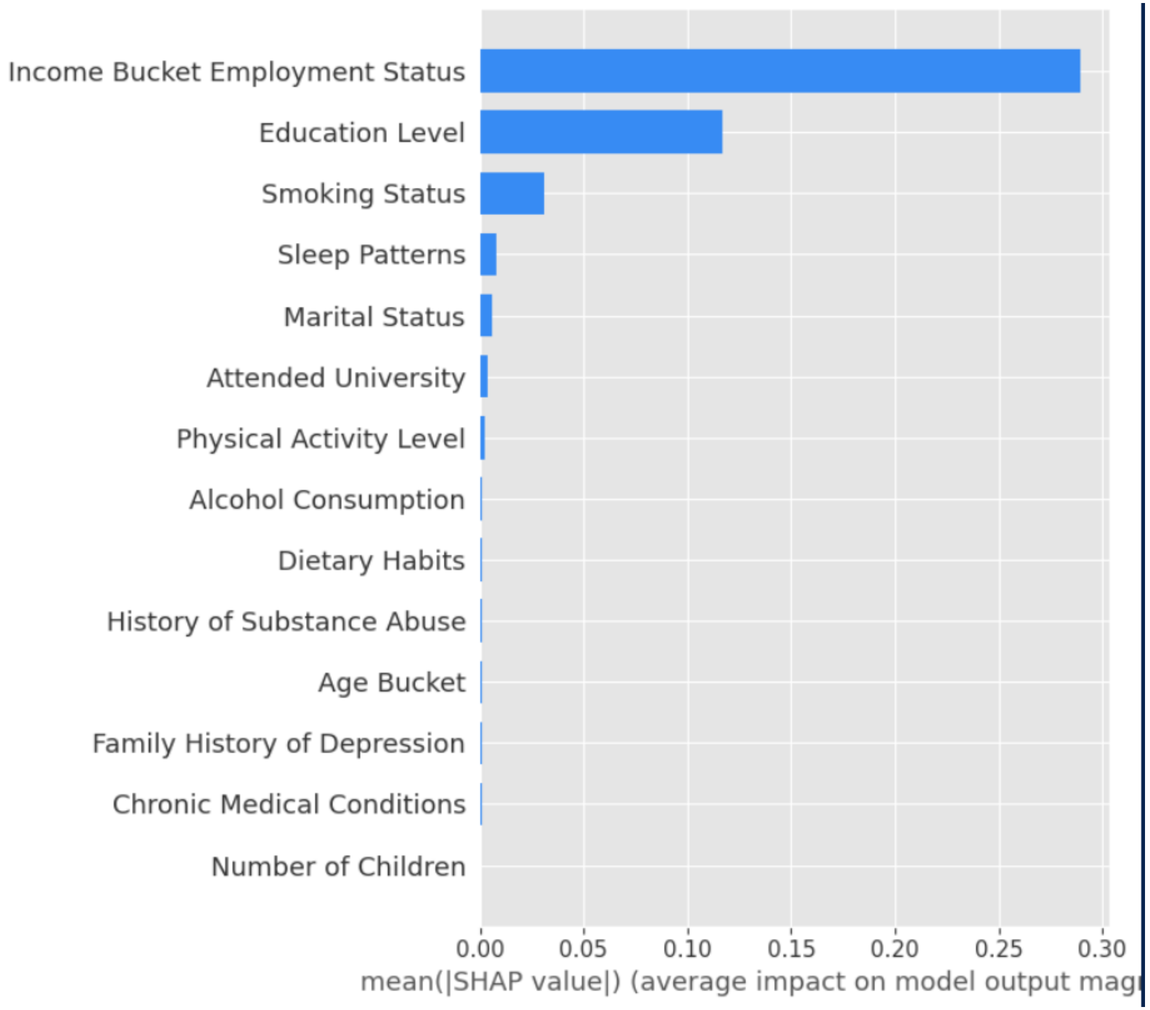
	Accuracy	Recall	Roc_Auc	Precision	F1 Score
Catboost_Weight_3	0.4776	0.8092	0.571	0.3464	0.4851





Modelling Results – Contributing Features

- Using SHAP analysis on the model we can see that the combined income and employment status bucket has the greatest impact.
- Education Level and Smoking Status have the next most impact.
- Less income and being unemployed has a correlation with the increase in likelihood of depression.
- Smoking increase correlates with increase in likelihood of Depression



Business Impact & Next Steps

Why This Matters?

- Helps **identify at-risk individuals early**, allowing for **interventions & support**.
- **Prioritising recall** means we minimize missed cases, even if some predictions are incorrect.

Next Steps

- Improve **explainability** (so non-experts can understand why the model flags individuals).
- **Explore combining more correlated features** (e.g., education level & income).
- Consider **real-world deployment** and ongoing monitoring to **reduce bias**.
- Obtain **more data** to help improve the model accuracy.