



Exploring the Depression statistics

Pod 11



Meet our members!



Peilu Tu



Lunark Vu



Gloria Achom



Alfred Okorochoa



Alondra Franco



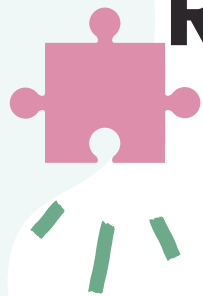
Courtney Wright





Our data:

Name	Age	Marital Status	Education Level	Number of Children	Smoking Status	Physical Activity Level	Employment Status	
Christine Barker	31	Married	Bachelor's Degree	2	Non-smoker	Active	Unemployed	
Jacqueline Lewis	55	Married	High School	1	Non-smoker	Sedentary	Employed	
Shannon Church	78	Widowed	Master's Degree	1	Non-smoker	Sedentary	Employed	
Charles Jordan	58	Divorced	Master's Degree	3	Non-smoker	Moderate	Unemployed	
Income	Alcohol Consumption	Dietary Habits	Sleep Patterns	History of Mental Illness	History of Substance Abuse	Family History of Depression	Chronic Medical Conditions	
26265.67	Moderate	Moderate	Fair	Yes	No	Yes	Yes	
42710.36	High	Unhealthy	Fair	Yes	No	No	Yes	
125332.79	Low	Unhealthy	Good	No	No	Yes	No	
9992.78	Moderate	Moderate	Poor	No	No	No	No	



Recap: Last Week's Decision to Use Smote

Before:

Logistic Regression			Random Forest Classifier		
	precision	recall		precision	recall
0	0.70	1.00	0	0.74	0.67
1	0.00	0.00	1	0.38	0.47

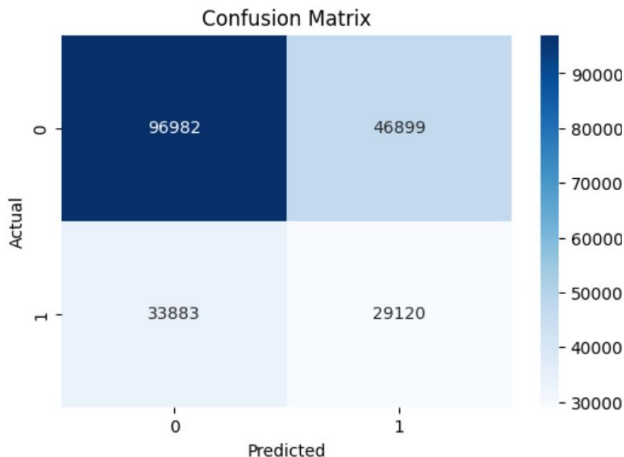
After:

Logistic Regression			Random Forest Classifier		
	precision	recall		precision	recall
No	0.63	0.63	No	0.64	0.64
Yes	0.63	0.63	Yes	0.64	0.65

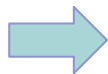
Our Focus Now: Random Forest Classifier

Parameters:

```
RandomForestClassifier(  
    • n_estimators=100,  
    • max_depth=5,  
    • random_state=42,  
    • n_jobs=-1  
)
```



Results:



	Method	Train Accuracy	Test Accuracy
0	Random Forest	0.696342	0.695467

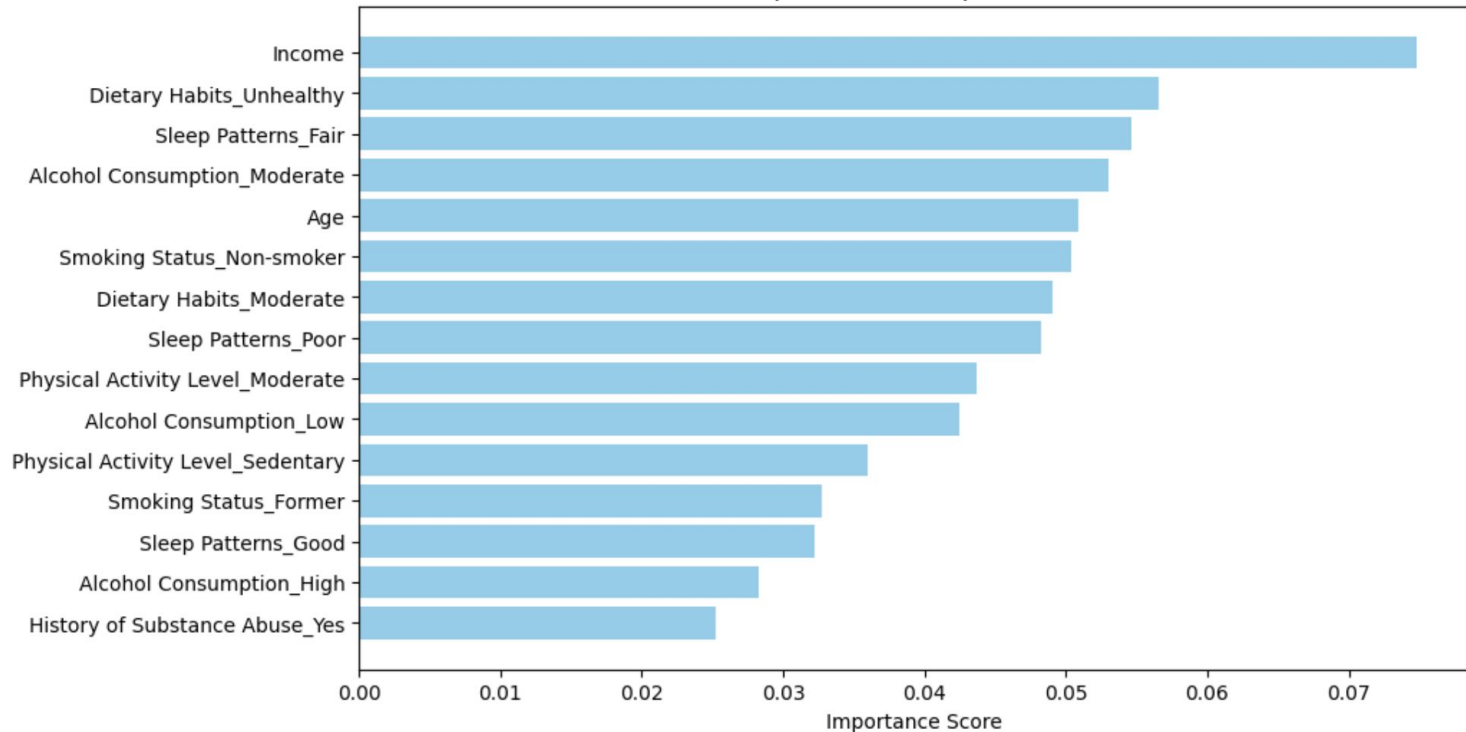
Feature Selection in Random Forest Classifier



Idea: remove
least important
features for better
performance



Top 15 Feature Importances





Feature Selection in Random Forest Classifier Results



```
=== Random Forest (Optimized) ===
```

```
Test Accuracy: 0.6483
```

```
Classification Report:
```

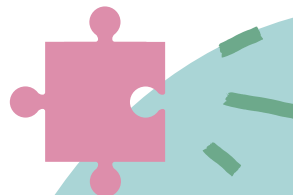
	precision	recall	f1-score	support
0	0.72	0.81	0.76	71986
1	0.39	0.27	0.32	31456
accuracy			0.65	103442
macro avg	0.55	0.54	0.54	103442
weighted avg	0.62	0.65	0.63	103442



Results

- Dropped 3 least important features:

Sleep Patterns(Good),
Alcohol
Consumption(High),
Substance Abuse



Generating Interaction Features in Random Forest Classifier



`PolynomialFeatures(degree=2, interaction_only=True, include_bias=False)`

Before:

```
Train Accuracy: 0.6927605958593609
Test Accuracy: 0.6757494512821938
Classification Report for Test Data:
```

	precision	recall
0	0.66	0.73
1	0.70	0.62
accuracy		
macro avg	0.68	0.68
weighted avg	0.68	0.68

After:

```
Train Accuracy: 0.8693812030191242
Test Accuracy: 0.692949555006066
Classification Report for Test Data:
```

	precision	recall
0	0.67	0.77
1	0.73	0.61
accuracy		
macro avg	0.70	0.69
weighted avg	0.70	0.69



Binning Features in Random Forest Classifier

Before:

```
Train Accuracy: 0.6927605958593609
Test Accuracy: 0.6757494512821938
Classification Report for Test Data:
              precision    recall
0             0.66         0.73
1             0.70         0.62

accuracy
macro avg         0.68         0.68
weighted avg      0.68         0.68
```

After:

```
Train Accuracy: 0.7268989149057681
Test Accuracy: 0.6740755146834106
Classification Report for Test Data:
              precision    recall
0             0.65         0.76
1             0.71         0.59

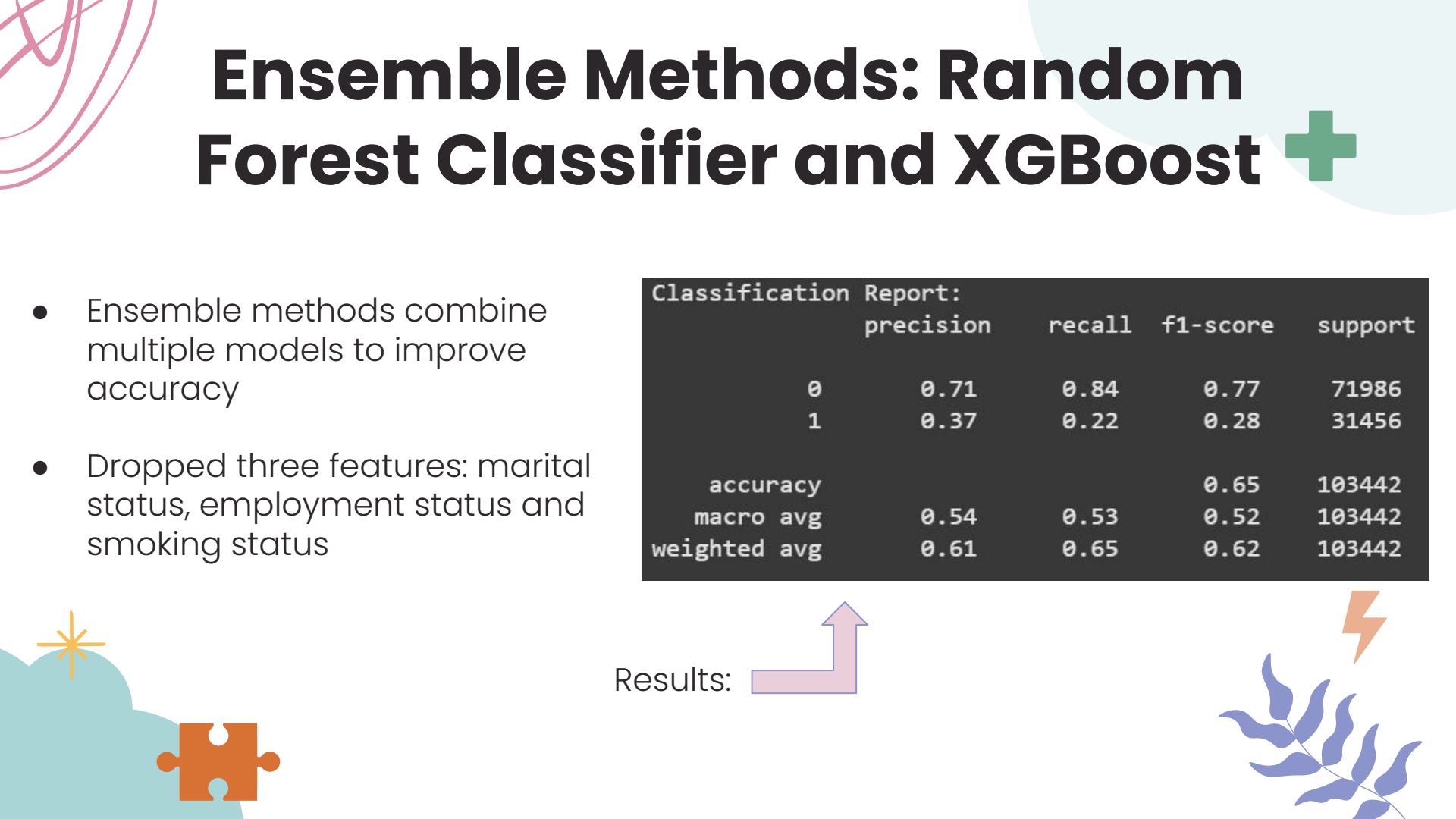
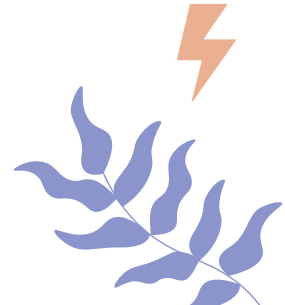
accuracy
macro avg         0.68         0.67
weighted avg      0.68         0.67
```

Ensemble Methods: Random Forest Classifier and XGBoost

- Ensemble methods combine multiple models to improve accuracy
- Dropped three features: marital status, employment status and smoking status

Classification Report:		precision	recall	f1-score	support
	0	0.71	0.84	0.77	71986
	1	0.37	0.22	0.28	31456
accuracy				0.65	103442
macro avg		0.54	0.53	0.52	103442
weighted avg		0.61	0.65	0.62	103442

Results:



Summary

- Random Forest Classifier is our best model
- SMOTE balanced our precision and recall
- Feature Selection improved our model
- Generating Interaction Features improved our model but caused overfitting
- Ensemble methods also caused overfitting

Timeline...for now



Week 2 (Jan 20–Jan 26): Model Development:

- Train three classification models
- Evaluate initial performance and address issues (e.g., class imbalance)

Week 3 (Jan 27–Feb 2): Refinement:

- Compare models based on metrics like accuracy, precision, recall, and F1-score

Week 4 (Feb 3–Feb 9): More refinement:

- Attempt improvement techniques like feature selection, interaction features, hyperparameter tuning, ensemble methods
- Mitigate overfitting of models
- Try XGBoost and compare results

Week 5 (Feb 10–Feb 16): Interpretation:

- Visualize results (e.g., feature importance, confusion matrices).
- Draft conclusions about which model is best and why.

We thank you for your eyeballs

Any question?

Here's our linktree :3

<https://linktr.ee/mayihaveamediumespresso>

