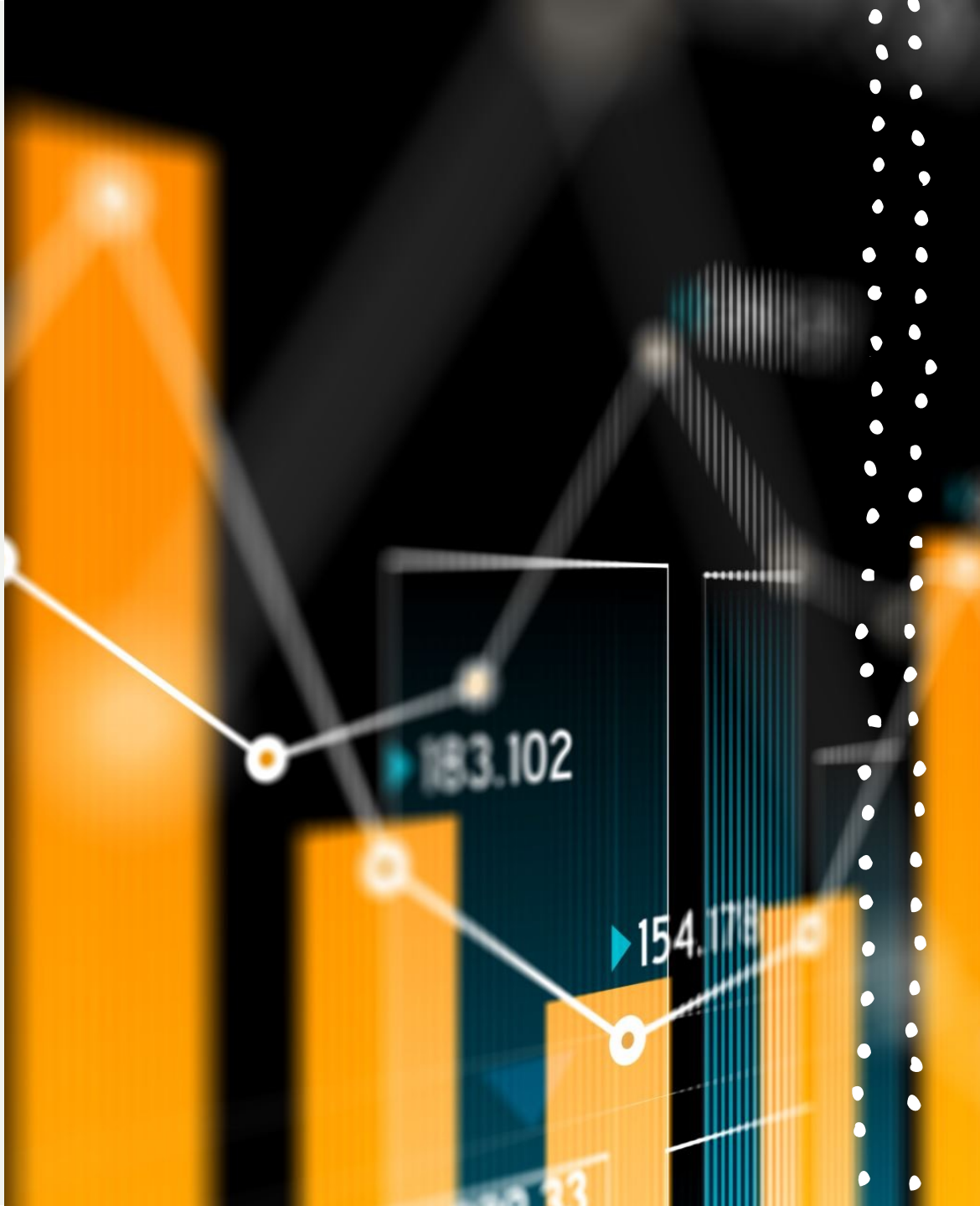

TITANIC "THE UNSINKABLE SHIP"

PROBLEM STATEMENT

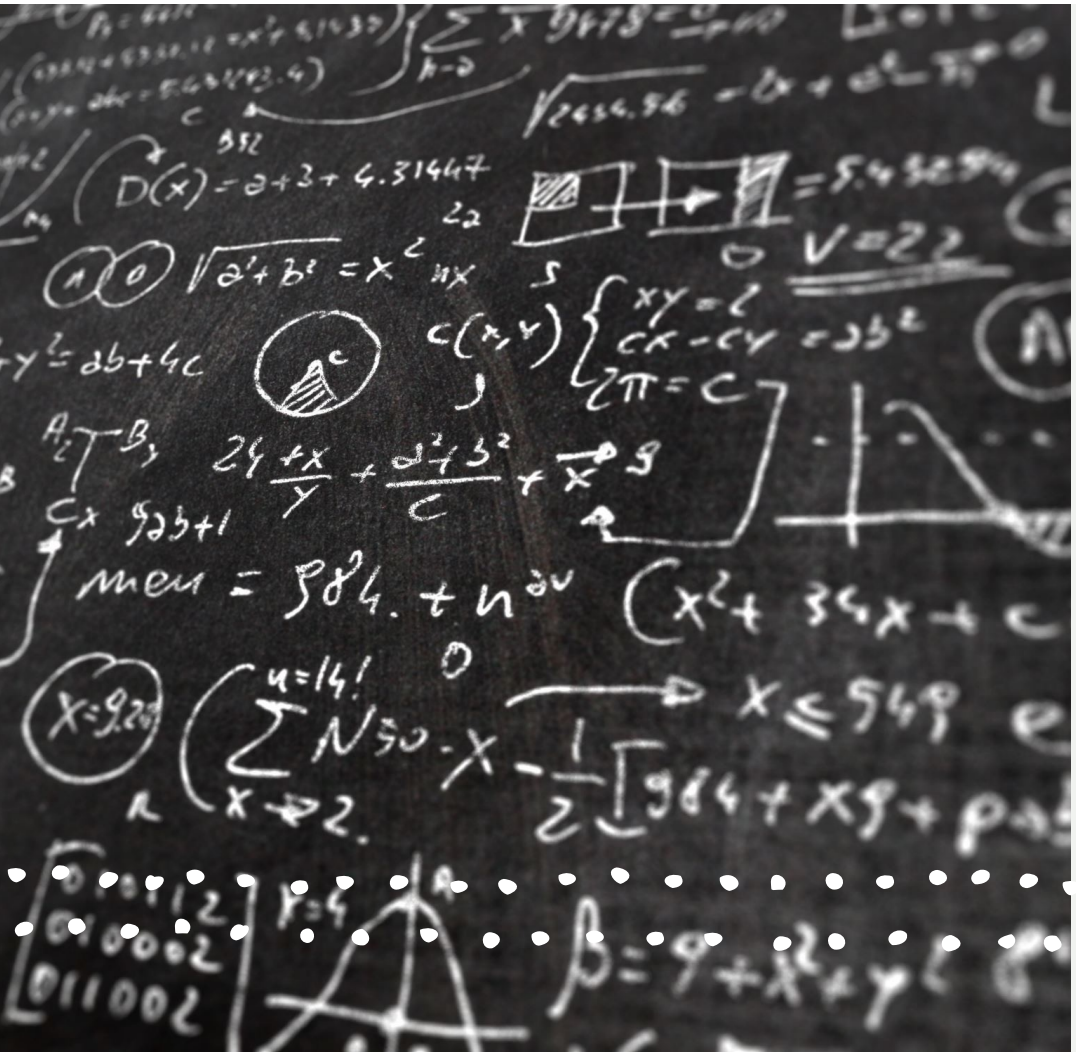
- The Titanic was a great and unexpected disaster in the early 1900's. The "unsinkable" ship struck an iceberg causing fatal damage to the integrity of the ship.
- This model is to accurately predict the survival of passengers aboard the ship the early morning of April 15th, 1912.

BUSINESS VALUE

- Upskilling of entry level Data Scientists by tackling a binary classification problem which can then be used in real world applications.
- Encouraging creative / "outside the box" thinking to feature engineer and improve model performance.



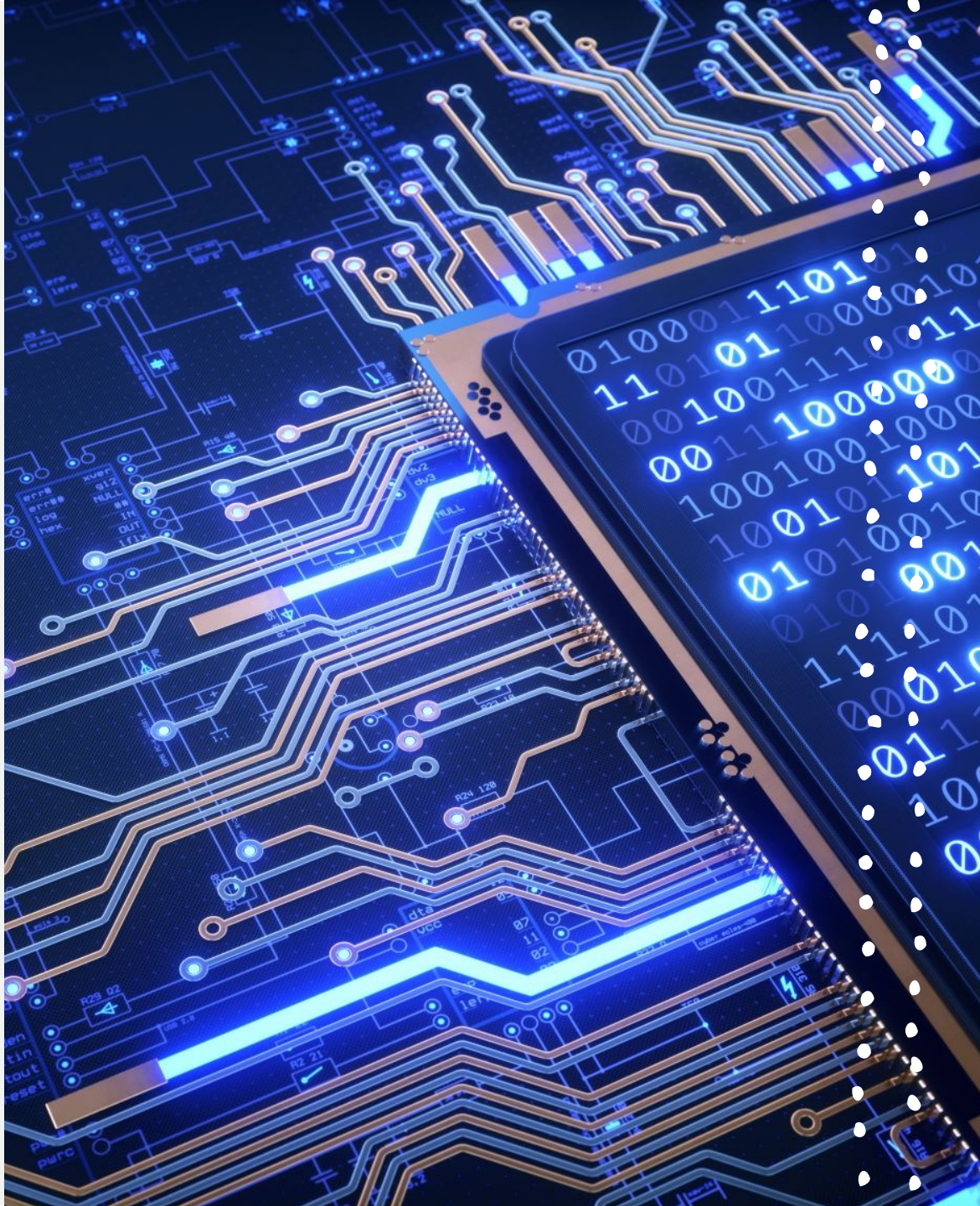
METHODOLOGY



- Analysis of Titanic data to understand key features of surviving passengers.
- Binary classification using XGBoost.
- Feature engineering to improve model performance and accuracy.
- Bayesian Optimization of tuning parameters to target well fit models with high model scores.

Feature Engineering

- New features:
  - **Title of an individual**
  - **Family Size**
  - **Fare cost per Person**
  - **Parent: Mother / Father**
  - **Family Demographics**
- Feature Bin:
  - **Family Size**
  - **Age**

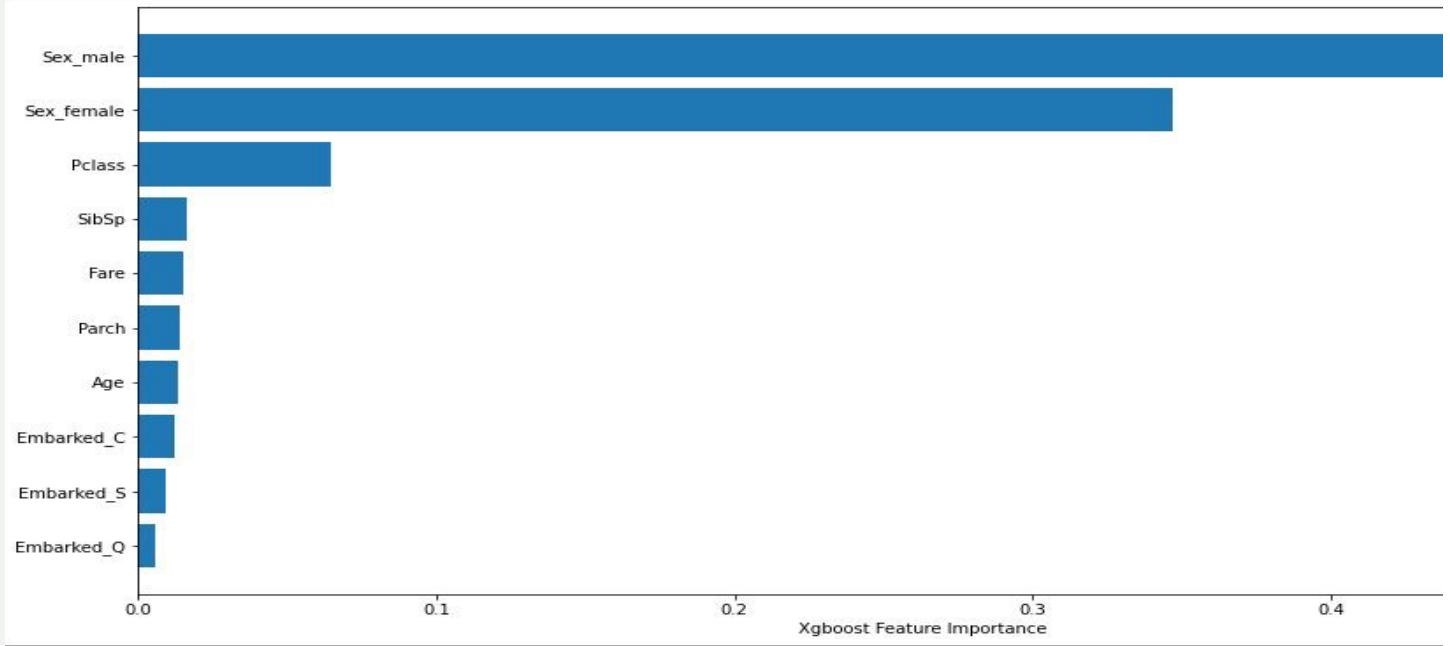


BASELINE MODEL

	precision	recall	f1-score	support
0	0.89	0.90	0.90	266
1	0.82	0.80	0.81	152
accuracy			0.87	418
macro avg	0.86	0.85	0.85	418
weighted avg	0.87	0.87	0.87	418

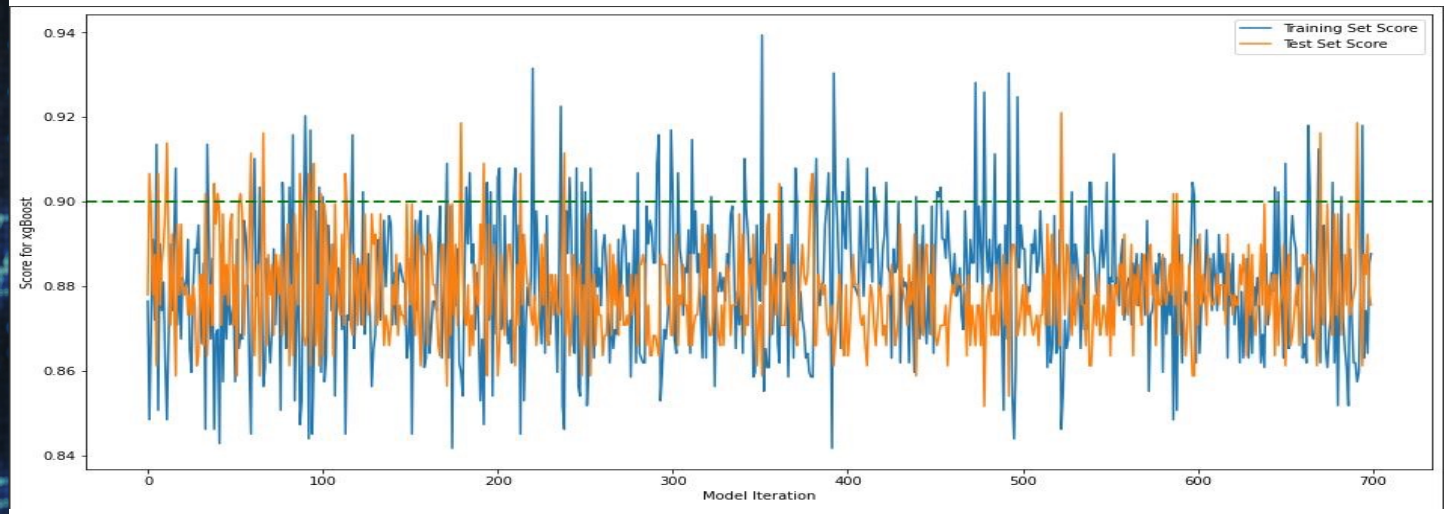
- BASELINE MODEL RESULTS [Overfit]
  - Training Score: 94.2%
  - Test Score: 86.6%
- Feature Importance:
  - Gender: Sex_male / Sex_female

```
Training set score for xgBoost: 0.942761
Testing set score for xgBoost: 0.866029
```

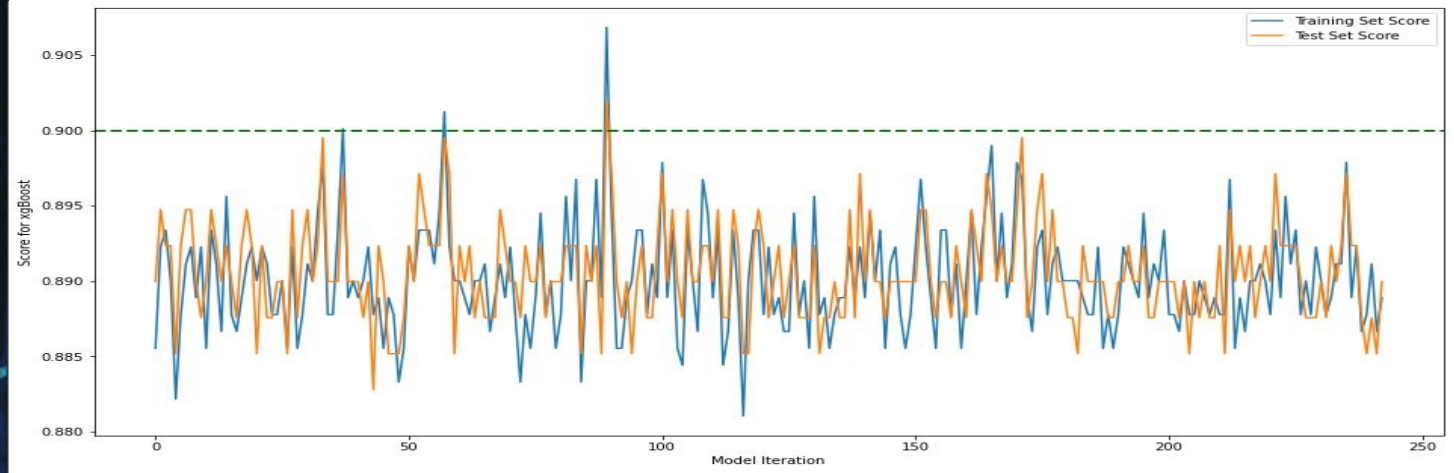


BAYESIAN OPTIMIZATION

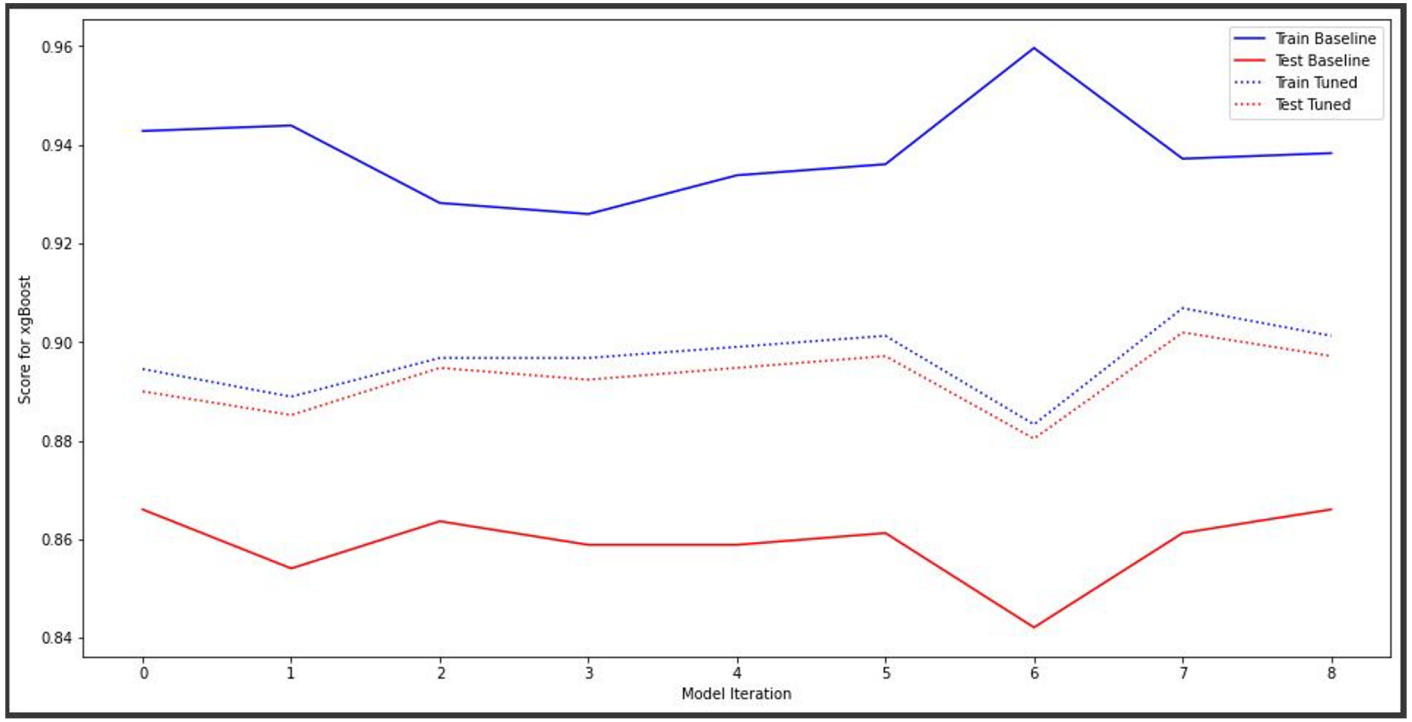
- Iteration of hyperparameter tuning to target ideal model performance.



- Targeting hyperparameters that have an accuracy < 0.005



BASELINE / TUNED MODEL EVOLUTION



- Feature Engineering Impact on Model Performance:

  Bin of Family Size – Negative Impact 1.7%

  Family Gender Count – Positive Impact 2.1%

SUMMARY:

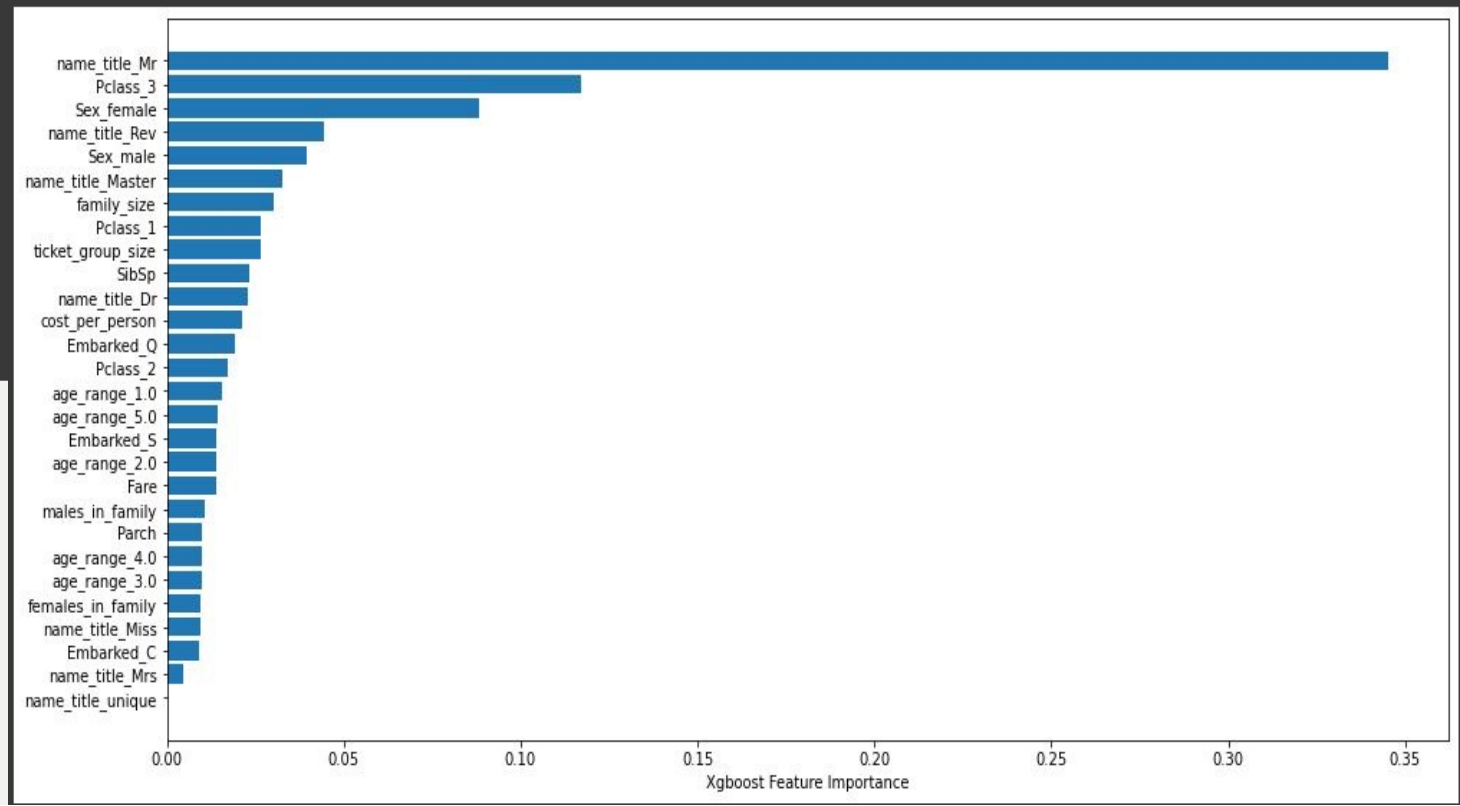
BEST MODEL

	precision	recall	f1-score	support
0	0.93	0.91	0.92	266
1	0.85	0.88	0.87	152
accuracy			0.90	418
macro avg	0.89	0.90	0.89	418
weighted avg	0.90	0.90	0.90	418

- BEST MODEL RESULTS
  - Training Score: 90.68%
  - Test Score: 90.19%
- Feature Importance:
  - Gender: name_title_Mr / Pclass_3

  Engineered features such as “name_title” and “family_size” ranked high in feature importance, yielding 50% of the top 10 important features.

```
Training set score for xgBoost: 0.906846
Testing set score for xgBoost: 0.901914
```



FUTURE CONSIDERATIONS

- Increase complexity of Bayesian optimization to expand the ”universe” of hyperparameters being tuned and their limits
- Automating column selection to iterate all combinations of columns used and columns to one-hot encode
- Utilizing GPU power to support more advanced optimization practices




Thank you