YUG PATEL (002842453)

Module 4 Leveraging AutoML Assignment

EAI 6020: AI System Technologies

03/22/25

Predictive Modeling Using AutoML: Titanic Survival Analysis

Introduction

This report presents an analysis conducted on the Titanic dataset to build a predictive model using Automated Machine Learning (AutoML). The primary goal was to predict passenger survival, providing actionable insights relevant to emergency preparedness and strategic planning.

Dataset Selection

The Titanic dataset was selected due to its wide recognition and straightforward structure, which makes it ideal for predictive modeling and educational purposes. It contains variables including passenger class (pclass), gender (sex), age (age), number of siblings/spouses aboard (sibsp), number of parents/children aboard (parch), fare paid (fare), and port of embarkation (embarked). The target variable is survived, indicating whether a passenger survived (1) or not (0). The dataset was retrieved from a publicly accessible source on GitHub (Datascience Dojo).

Economic Viability of AI Solution Variables

Each selected variable has economic and operational implications, particularly in fields like emergency response, insurance, healthcare, and risk management. For example, age and gender are critical determinants in survival prediction, influencing evacuation priorities and resource allocation in emergencies. Economic viability emerges from improved accuracy in predicting outcomes, enabling more efficient use of resources and better strategic planning. Variables such as fare reflect economic status, potentially correlating with survival rates and allowing for socio-economic analyses relevant to insurance and risk assessment sectors.

```
<class 'pandas.core.frame.DataFrame'>
Index: 712 entries, 0 to 890
Data columns (total 8 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   survived  712 non-null    int64
 1   pclass    712 non-null    int64
 2   sex       712 non-null    object
 3   age       712 non-null    float64
 4   sibsp     712 non-null    int64
 5   parch     712 non-null    int64
 6   fare      712 non-null    float64
 7   embarked  712 non-null    object
dtypes: float64(2), int64(4), object(2)
memory usage: 50.1+ KB
```

Training and Evaluation using Precision-Recall Curve

AutoML, specifically through PyCaret, was employed for model training. PyCaret automatically compares multiple machine learning algorithms, handles feature engineering, and optimizes model selection. The dataset was split into training and testing sets. PyCaret identified the Light Gradient Boosting Machine (LightGBM) as the optimal model, achieving an accuracy of approximately 81.31%, with notable metrics including an AUC of 0.8529, precision of 0.7924, and recall of 0.7360.

Model performance evaluation leveraged the precision-recall curve, a robust tool especially suited for binary classification with class imbalances, such as survival prediction. Unlike the ROC curve, precision-recall curves directly address class imbalance by emphasizing false positives and false negatives, which are critical in survival scenarios. Precision reflects the proportion of true positives among all positive predictions, whereas recall represents the proportion of actual positives correctly identified.

```
# Compare multiple models and select the best one
best_model = compare_models()
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.8131	0.8529	0.7360	0.7924	0.7575	0.6065	0.6126	0.1610
rf	Random Forest Classifier	0.8111	0.8530	0.7310	0.7919	0.7497	0.6002	0.6092	0.3500
xgboost	Extreme Gradient Boosting	0.8091	0.8388	0.7310	0.7871	0.7516	0.5977	0.6045	0.2200
gbc	Gradient Boosting Classifier	0.8070	0.8601	0.7310	0.7818	0.7486	0.5933	0.6003	0.1830
knn	K Neighbors Classifier	0.8051	0.8395	0.7557	0.7597	0.7548	0.5933	0.5965	0.0910
et	Extra Trees Classifier	0.7949	0.8465	0.7162	0.7584	0.7308	0.5667	0.5721	0.2340
dt	Decision Tree Classifier	0.7891	0.7800	0.7360	0.7406	0.7341	0.5599	0.5640	0.0740
ridge	Ridge Classifier	0.7890	0.8522	0.7412	0.7406	0.7376	0.5616	0.5649	0.0910
lda	Linear Discriminant Analysis	0.7890	0.8519	0.7412	0.7406	0.7376	0.5616	0.5649	0.0710
lr	Logistic Regression	0.7829	0.8516	0.7462	0.7267	0.7331	0.5507	0.5542	0.8140
ada	Ada Boost Classifier	0.7809	0.8369	0.7462	0.7253	0.7330	0.5478	0.5507	0.1550
qda	Quadratic Discriminant Analysis	0.7628	0.8104	0.7464	0.6979	0.7191	0.5150	0.5182	0.0730
nb	Naive Bayes	0.7549	0.8128	0.7317	0.6920	0.7076	0.4976	0.5025	0.0730
svm	SVM - Linear Kernel	0.7045	0.7421	0.7162	0.6285	0.6622	0.4038	0.4136	0.0750
dummy	Dummy Classifier	0.5964	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0720

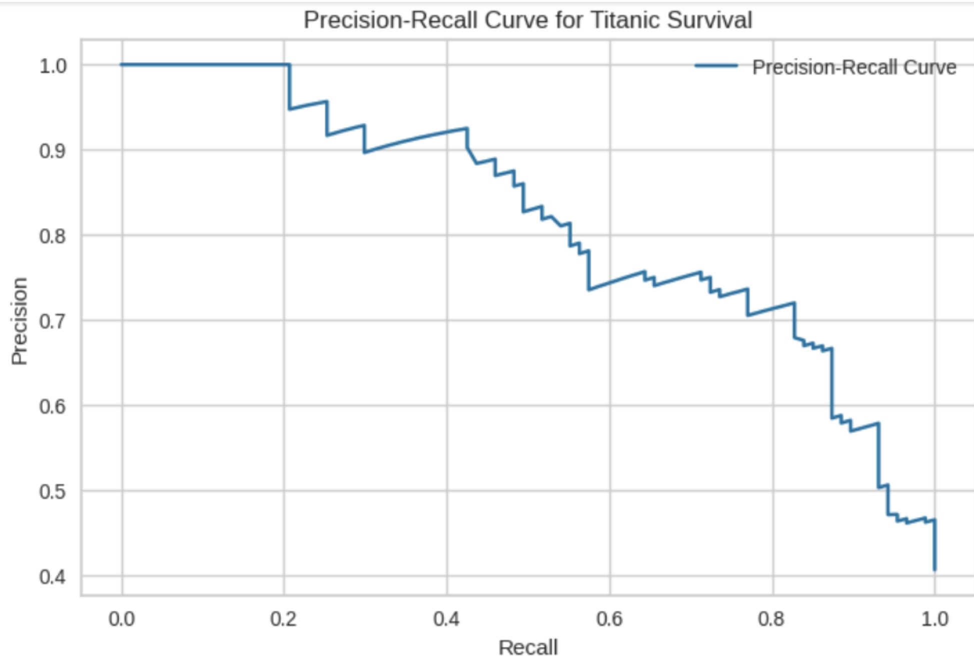
```
[ ] predictions = predict_model(best_model, raw_score=True)

    # Verify columns
    print(predictions.columns)
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Light Gradient Boosting Machine	0.7710	0.8520	0.6667	0.7436	0.7030	0.5176	0.5197

```
Index(['pclass', 'sex', 'age', 'sibsp', 'parch', 'fare', 'embarked',
       'survived', 'prediction_label', 'prediction_score_0',
       'prediction_score_1'],
      dtype='object')
```

Setting an Appropriate Score Threshold

Choosing an appropriate probability threshold was crucial for operational decision-making. Analysis of the precision-recall curve indicated that a threshold of 0.6 provided the most balanced trade-off, with acceptable precision and recall values (precision: 0.74, recall: 0.57), leading to an F1 score of 0.65. This threshold effectively balanced the cost of false positives against false negatives, optimizing both economic and operational effectiveness.



	precision	recall	f1-score	support
0	0.75	0.86	0.80	127
1	0.74	0.57	0.65	87
accuracy			0.74	214
macro avg	0.74	0.72	0.72	214
weighted avg	0.74	0.74	0.74	214

Justification for Targeting Survival

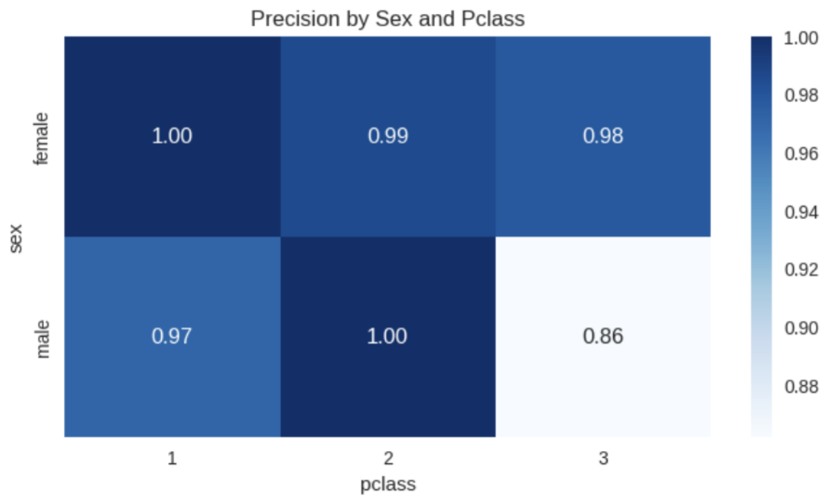
The decision to target the survival outcome directly aligns with practical applications in emergency response planning, disaster management, and insurance risk assessment. Predicting survival provides clear economic and humanitarian value, enhancing decision-making processes during crises. Effective prediction models can inform strategies that improve resource allocation, saving lives and reducing economic losses.

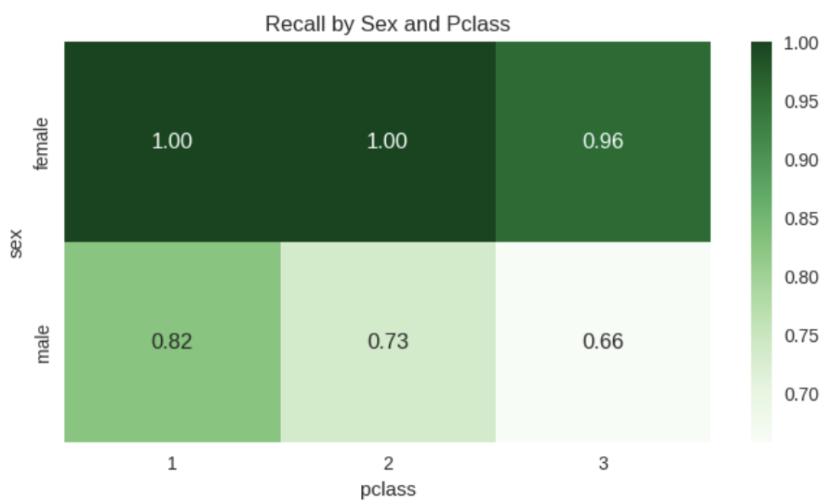
Precision Heatmap Insights

Precision measures how many of the predicted survivors among passengers actually survived.

- Female passengers across all classes (1st, 2nd, and 3rd) exhibit remarkably high precision (ranging from 0.98 to 1.00), indicating that the model is very dependable in forecasting survival for women.
- Males in both 1st and 2nd class also display high precision (0.97 and 1.00), but this figure decreases to 0.86 in 3rd class, indicating a higher rate of false positives for lower-class male travelers.

0	Light Gradient Boosting Machine	0.9537	0.9945	0.9097	0.9740	0.9408	0.9028	0.9042





Recall Heatmap Insights

Recall measures how many true survivors were accurately identified by the model.

- The recall for females remains consistently high across all categories (0.96–1.00), showing that the model performs exceptionally well in recognizing actual female survivors.

- In contrast, male recall is significantly lower, especially for the 3rd class (0.66), followed by the 2nd class (0.73), and then the 1st class (0.82). This suggests that the model tends to overlook a greater number of male survivors, particularly in the 3rd class.

Lessons Learned and Future Implications

This exercise provided several valuable lessons. Firstly, the importance of data preprocessing and cleaning cannot be overstated, as these significantly impact model performance. Secondly, AutoML considerably streamlines model development, allowing analysts to focus more on interpretation and application rather than manual hyperparameter tuning. Additionally, precision-recall curves offer more relevant insights in scenarios of class imbalance, guiding better-informed operational decisions.

Moving forward, future projects will prioritize incorporating AutoML to enhance productivity and improve predictive accuracy. Additionally, a structured and business-aligned approach will be consistently adopted in setting thresholds for classification tasks, emphasizing precision and recall metrics based on business or operational implications.

Conclusion

Using AutoML to analyze Titanic survival data demonstrated the effectiveness of automated processes in predictive modeling. By setting an optimal threshold informed by a precision-recall analysis, this approach offers significant potential for application across various economically and operationally significant domains. Embracing these methodologies ensures not only efficient analysis but also impactful, data-driven decisions.

References

1. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
2. Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., & Moore, J. H. (2018). Data-driven advice for applying machine learning to bioinformatics problems. Pacific Symposium on Biocomputing 2018, 23, 192-203.
3. Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84-90. https://doi.org/10.1016/j.inffus.2021.11.011