**Titanic Passenger Survival Prediction - Machine Learning (Dataiku)**

**Project Report**

This project utilizes machine learning methods to predict the survival probability of Titanic passengers. The model was developed using the Dataiku DSS platform and includes data preprocessing, model training, and evaluation phases.  
This project was carried out using the **Random Forest** model to predict the survival probability of Titanic passengers.  
Our objective is to examine the performance of the **Random Forest** model in this type of problem by applying data preprocessing, model training, and evaluation processes.  
The model was successfully trained, and satisfactory results were obtained.

**Dataset**

The **Kaggle Titanic dataset** was used in this project. This dataset contains information about passengers such as age, gender, ticket class, number of family members onboard, and ticket fare.

The dataset includes the following files:

* **train.csv** → Labeled data used for model training (Survived: 0 = Did not survive, 1 = Survived)
* **test.csv** → Unlabeled dataset for predictions
* **predictions.csv** → Prediction results generated by the model

**Evaluation Metrics**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Doğruluk (Accuracy)** |  |  | | --- | |  | | |  | | --- | | 0.846 |  |  | | --- | |  | |
| |  | | --- | | **Kesinlik (Precision)** |  |  | | --- | |  | | |  | | --- | | 0.769 |  |  | | --- | |  | |
| |  | | --- | | **Duyarlılık (Recall)** |  |  | | --- | |  | | |  | | --- | | 0.877 |  |  | | --- | |  | |
| |  |  |  | | --- | --- | --- | | |  | | --- | | F1-Score |  |  | | --- | |  | |  |  | | --- | |  | | 0.820 |
| |  |  |  | | --- | --- | --- | | |  | | --- | | Cost Matrix Gain |  |  | | --- | |  | |  |  | | --- | |  | | |  | | --- | | 0.318 |  |  | | --- | |  | |
| |  | | --- | | **ROC AUC** Score |  |  | | --- | |  | | |  | | --- | | 0.889 |  |  | | --- | |  | |
| |  |  |  | | --- | --- | --- | | |  | | --- | | Lift Score |  |  | | --- | |  | |  |  | | --- | |  | | |  | | --- | | 2.025 |  |  | | --- | |  | |
| |  |  |  | | --- | --- | --- | | |  | | --- | | Average Precision |  |  | | --- | |  | |  |  | | --- | |  | | |  | | --- | | 0.887 |  |  | | --- | |  | |
| |  | | --- | | **Log Kayıp (Log Loss)** |  |  | | --- | |  | | |  | | --- | | 0.391 |  |  | | --- | |  | |
| |  |  |  | | --- | --- | --- | | |  | | --- | | Calibration Loss |  |  | | --- | |  | |  |  | | --- | |  | | 0.128 |

### ****Metric Interpretations****

* **Accuracy (0.846):**  
  → Indicates that **84.6%** of the model’s predictions were correct. For a dataset like Titanic, which is not highly imbalanced, this is a good score.
* **Precision (0.769):**  
  → Shows that **76.9%** of the model's positive survival predictions were actually correct. This means the model may be prone to **false positives**.
* **Recall (0.877):**  
  → The model correctly predicted **87.7%** of actual survivors. A high recall is beneficial as it reduces the risk of missing actual survivors. However, since precision is lower, there may be some **false positives**.
* **F1-Score (0.820):**  
  → Represents the balance between **precision and recall**. The model’s overall performance is **decent**, but the lower precision indicates that false positives might be an issue.
* **Cost Matrix Gain (0.318):**  
  → Reflects the impact of incorrect predictions. This is **not a very high value**, indicating that incorrect predictions somewhat reduce the model's effectiveness.
* **ROC AUC Score (0.889):**  
  → Demonstrates the model’s ability to **distinguish between classes**. A value of **0.889** suggests a **highly successful** model. A perfect model would score **1.0**, meaning this model performs well.
* **Lift Score (2.025):**  
  → Measures how much better the model is compared to random guessing. A score of **2.025** suggests that the model is **twice as effective** as random guessing.
* **Average Precision (0.887):**  
  → Evaluates the model’s **overall prediction accuracy**. A score of **0.887** is **very good**.
* **Log Loss (0.391):**  
  → Lower values are better, as they indicate **less uncertainty** in predictions. **0.391 is not too high**, meaning the model is **performing well**.
* **Calibration Loss (0.128):**  
  → Measures how well the predicted probabilities align with actual class distributions. Since **0.128** is a **low value**, the model’s probability estimates are **fairly reliable**.

## ****General Comments:****

* The model has **high recall (0.877)**, meaning it is **unlikely to miss actual survivors**.
* **Precision (0.769) is a bit low**, suggesting that **false positives** might be an issue. The model may **incorrectly classify non-survivors as survivors**.
* **ROC AUC Score (0.889) and Lift Score (2.025) are very good**, indicating that the model effectively differentiates between survivors and non-survivors.
* **Log Loss (0.391) is relatively low**, meaning there is **not much uncertainty** in predictions.
* **Calibration Loss (0.128) is low**, so the model's probability estimates are **trustworthy**.

## ****Conclusion:****

The model is **generally successful**, but since **precision is relatively low**, it may **produce more false positives**.  
If **false positives are critical** (e.g., mistakenly predicting someone survived when they didn’t), the model can be adjusted to **increase precision**.  
However, in its **current state**, the model is **performing well**.