Dhirubhai Ambani University
Technology
Formerly DA-IICT

IT457 Cloud Computing

Assignment-Amazon SageMaker
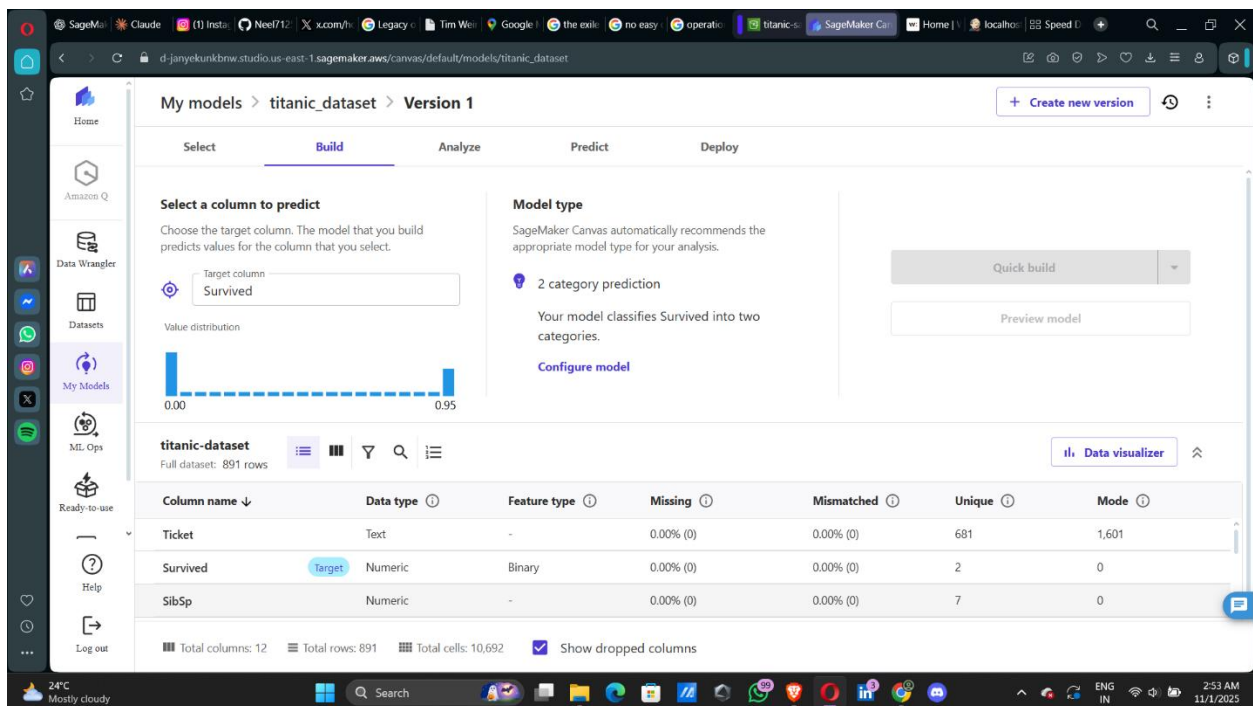
Group 5

| Name	Student ID
Sneh Joshi	202201048
Sahil Chaudhari	202201171
Kshitij Patel	202201232
Neel Patel	202201494

Model Analysis

Dataset Used : [Titanic Dataset](#)

Visualization Analysis

The Titanic dataset contains 891 rows and 12 columns, with **"Survived"** as the binary target variable (62% did not survive, 38% survived). It has **0% missing or mismatched values**, ensuring excellent data quality with no preprocessing required. Key features include **Ticket** (681 unique values, indicating shared bookings) and **SibSp** (7 unique values, mostly 0, meaning most traveled alone). The dataset's mix of **numeric, categorical, and text** features provides a strong foundation for building an accurate survival prediction model.

Feature Impact Analysis
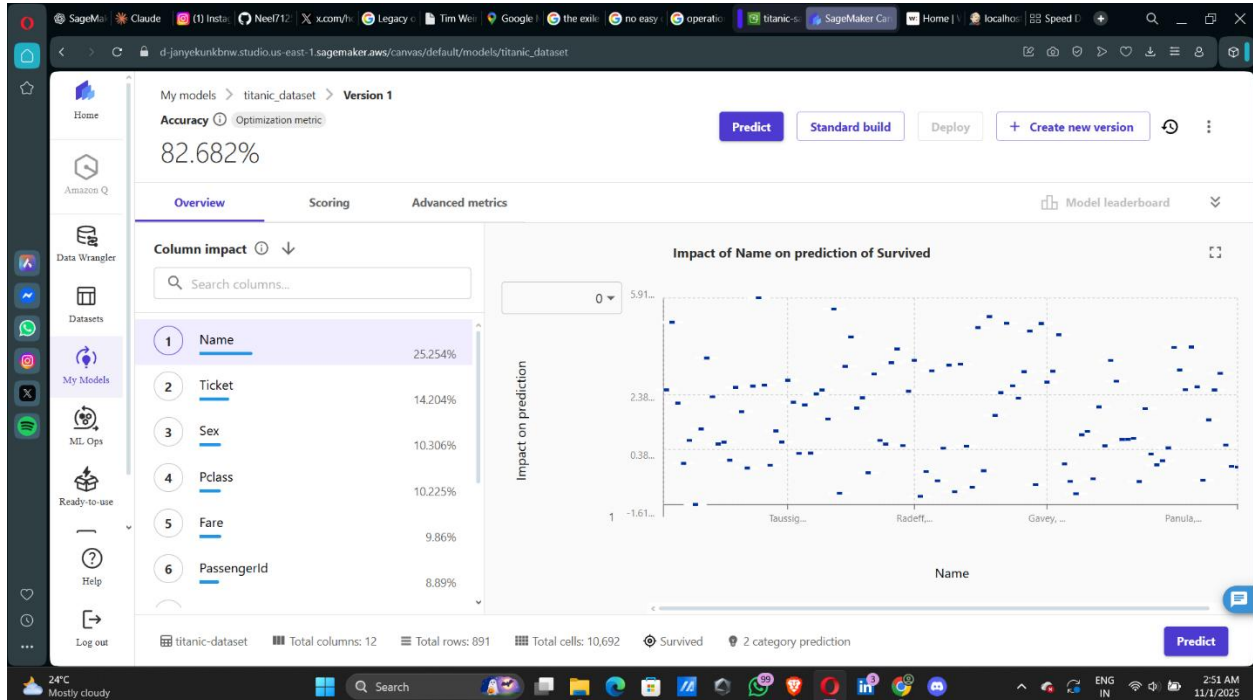
Name (25.25%) – The most influential feature, as titles within names (Mr., Mrs., Miss., etc.) reflected social status, gender, and age group. Titles like *Mrs.* and *Miss* strongly correlated with survival, while *Mr.* indicated lower chances.
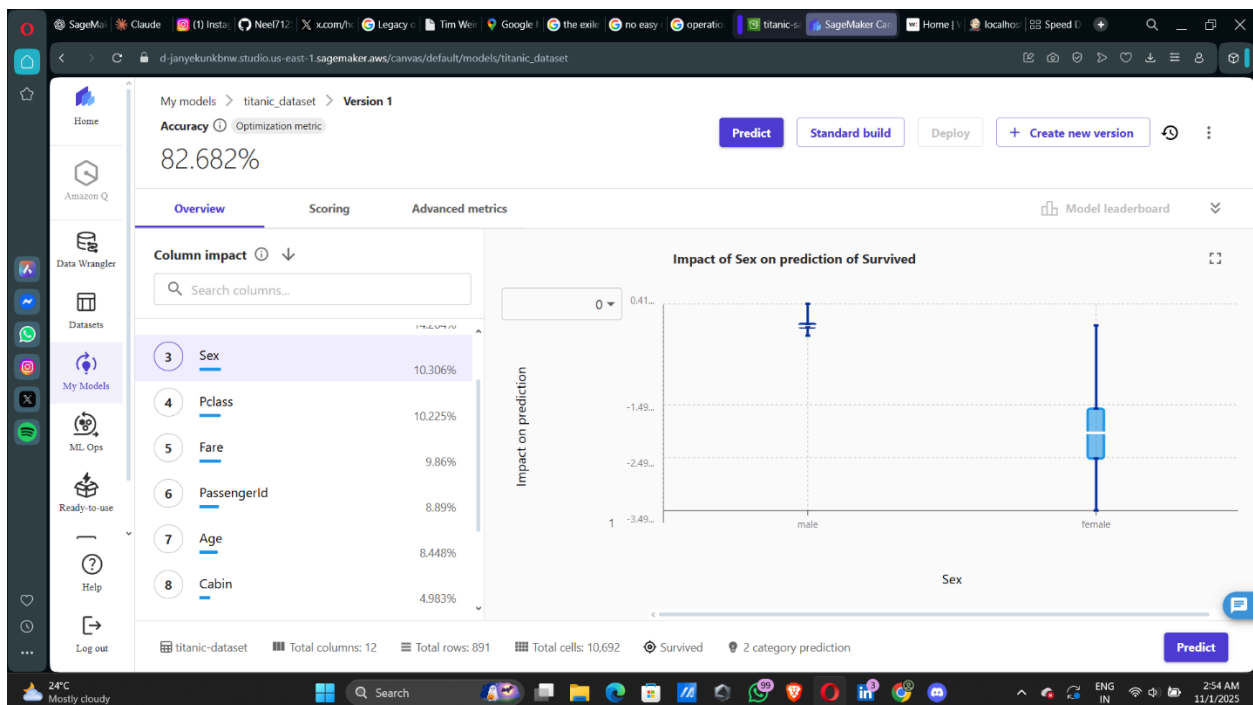
Sex (10.31%) – A major predictor showing women had much higher survival impact scores, aligning with the "women and children first" rule. Female passengers were prioritized for lifeboats, making gender a key survival determinant.
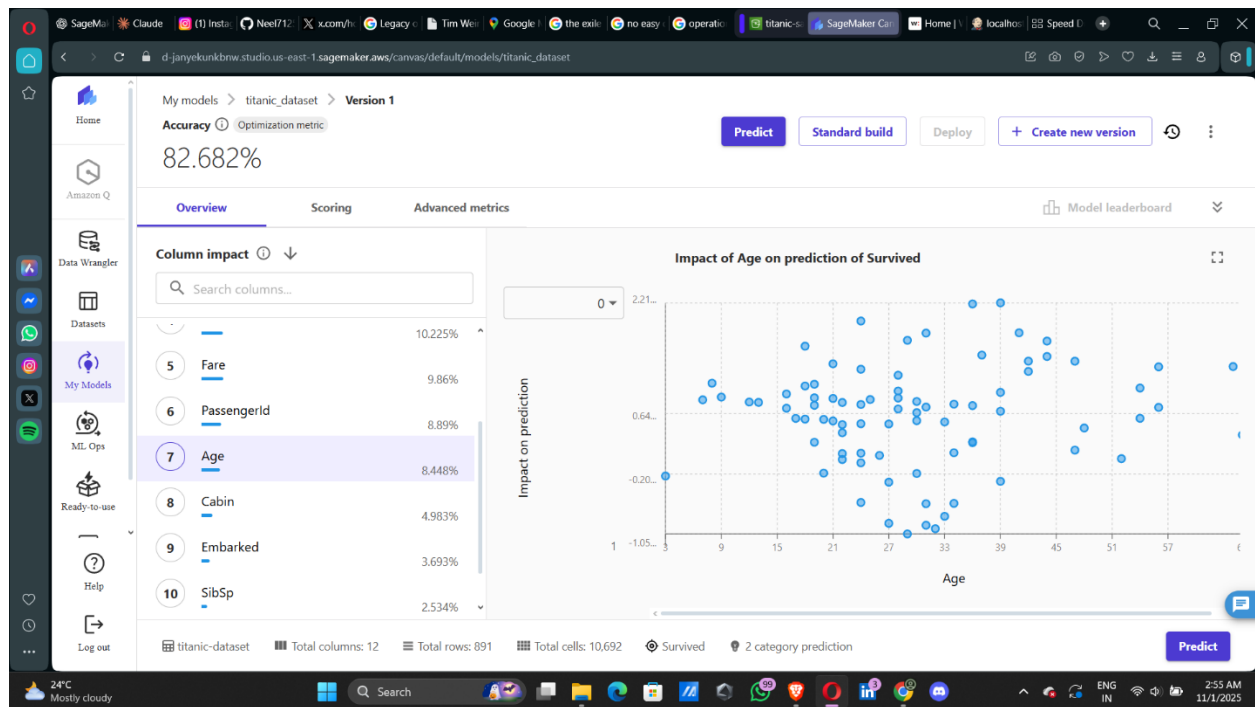
Age (8.45%) – Younger passengers (0–15) had higher survival impacts, supporting child priority during evacuation. For ages 20–40, survival varied widely based on other factors like gender and class, showing that age interacted with social and demographic variables.

Individual Passenger Prediction Analysis

The model analyzed a sample passenger, **Mr. Anthony Abbing**, a 24-year-old male in **third class (Pclass=3)** with no family aboard. It predicted a **98.2% chance of non-survival** and only **1.8% survival probability**. This outcome reflects several high-risk factors: being male, traveling in third class, being a young adult, and traveling alone—all historically associated with lower survival rates. The prediction tool effectively shows how combinations of **age, gender, class, and family status** influenced outcomes, offering insights into the social and demographic inequalities that shaped survival during the Titanic disaster.

Accuracy

The Sankey diagram in the Scoring tab visualizes the model's strong performance across **179 test predictions** with minimal misclassifications. For passengers predicted **not to survive (Class 0)**, most correctly map to actual non-survivors, while only a few represent false negatives. Similarly, for **predicted survivors (Class 1)**, most align with actual survivors, with few false positives. The model achieves **~85.6% accuracy for non-survival predictions** and correctly identifies **~86.4% of actual non-survivors**, reflecting balanced performance. The clear, minimal crossover in the diagram confirms that the model effectively distinguishes survivors from non-survivors without major bias toward either class.
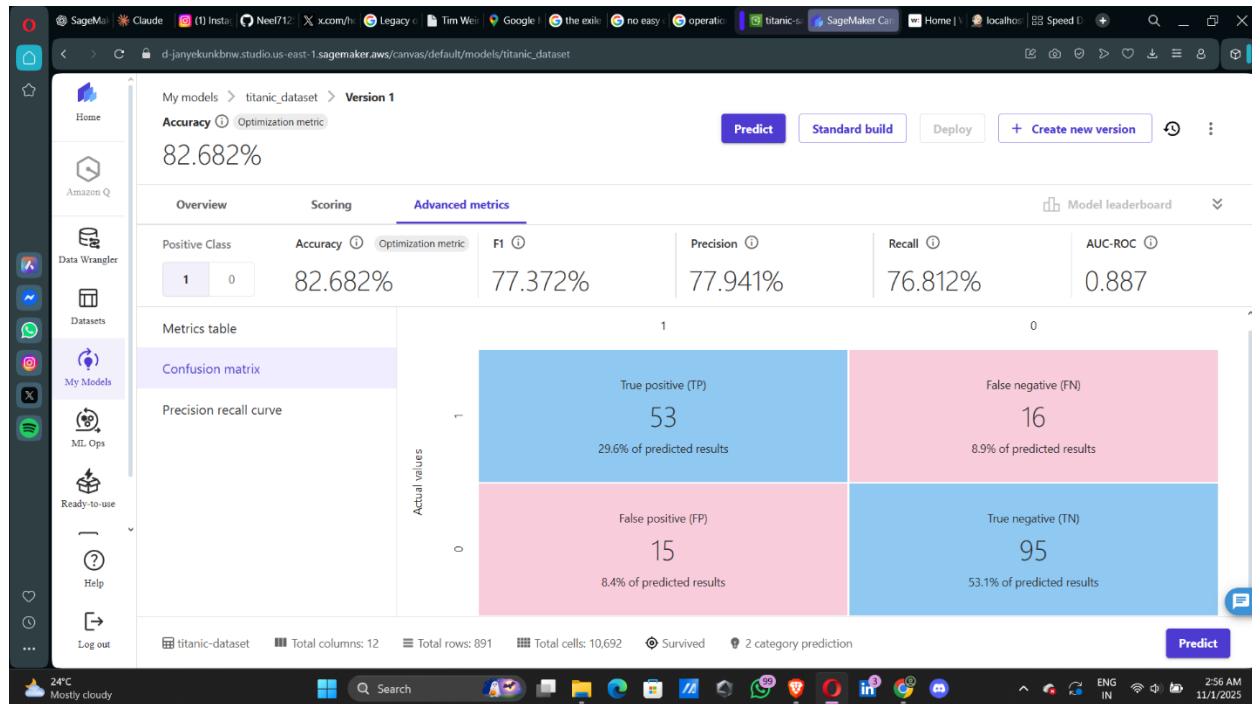
Advanced Metrics Analysis

Metrics table
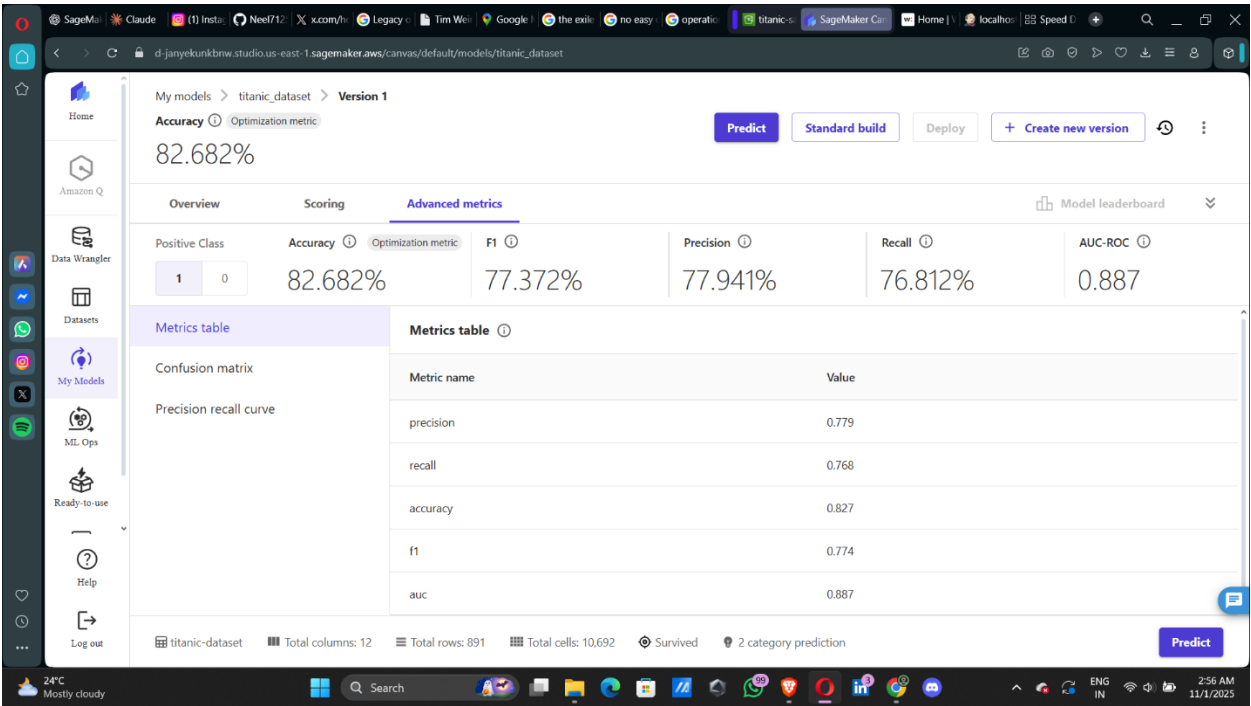
With **Precision = 0.779** and **Recall = 0.768**, the model accurately identifies most survivors while maintaining few false alarms. The **F1 Score (0.774)** reflects strong balance, and the **AUC (0.887)** confirms excellent separation between survival outcomes.

Confusion matrix

The confusion matrix breaks down the model's 179 test predictions into four categories: 53 True Positives (29.6% of predictions) where the model correctly identified survivors, 95 True Negatives (53.1%) where non-survivors were correctly predicted, 16 False Negatives (8.9%) representing survivors incorrectly classified as non-survivors, and 15 False Positives (8.4%) representing non-survivors incorrectly predicted to survive. The near-equal error rates (8.9% vs 8.4%) demonstrate the model's balanced performance without significant bias toward either class, while the dominant blue quadrants (TP and TN) visually confirm that correct predictions far outnumber errors.

Precision recall curve

The **Precision-Recall curve (AUPRC = 0.87)** shows that the model maintains **high precision (>0.8)** for recall values up to about 0.5, then gradually declines as recall increases. This indicates strong confidence in positive predictions with few false positives. The **AUPRC far above the baseline (≈0.38)** confirms the model's effectiveness in identifying real survivors. The smooth curve reflects **stable and reliable performance**, suggesting good generalization to unseen Titanic data.

d-janyekunkbnw.studio.us-east-1.**sagemaker.aws**/canvas/default/models/titanic_dataset

My models > titanic_dataset > **Version 1**

Accuracy ⓘ Optimization metric

82.682%

Predict | Standard build | Deploy | + Create new version

Overview | Scoring | **Advanced metrics** | Model leaderboard

Positive Class	Accuracy ⓘ Optimization metric	F1 ⓘ	Precision ⓘ	Recall ⓘ	AUC-ROC ⓘ
1 0	82.682%	77.372%	77.941%	76.812%	0.887

Metrics table

Confusion matrix

Precision recall curve

Precision vs. recall

● Precision-recall (AUPRC = 0.87)

🗀 titanic-dataset ⚏ Total columns: 12 ☰ Total rows: 891 ⊞ Total cells: 10,692 ◎ Survived 💡 2 category prediction

Predict