SkyCast: Predicting Space Mission Outcomes

An End-to-End Machine Learning Solution

Presentation By:- Kitiksha Chakrawarti

Roll Number:- 23F3001245

INTRODUCTION

Problem Statement

Build a predictive model that can accurately classify the outcome of space missions as either "Success" or "Failure" based on various features provided in the dataset.

Project Goal

To develop a robust and highly accurate classification model that generalizes well to unseen launch data.

Data Sources

- **train.csv**- The training set with 3500 rows and 9 columns
- **test.csv**- The test set
- **Data Source**- Kaggle competition dataset

Key Achievement-

Secured 8th position on the private leaderboard and 6th position after presentation and final evaluation

Data Cleaning and Preparation

Data Quality Control

Dropped features with >70% missing values and eliminating redundant columns.

Feature Isolation

Extracted key predictive features from compound columns

Data Encoding

Prepared target feature for the model by encoding into binary values

Process Control

Ensured a reproducible workflow by encapsulating all data preparation steps within a Scikit-learn Pipeline

1. Checking for missing values

```
df.isnull().sum()
```

2. Dropping repeated columns

```
(df['Unnamed: 0.1']==df['Unnamed: 0']).all()
```

```
np.True_
```

Detail
Voskhod | Cosmos 214

→

Rocket_type	Mission
Voskhod	Cosmos 214

Status Rocket
StatusRetired
StatusActive

→

Status Rocket
0
1

Advanced Feature Engineering

Temporal & Geographical Feature Creation

Derived predictive features for Time and Location by parsing complex original strings.

Dimensionality Reduction

Identified highly correlated features and reduced feature complexity by grouping low-frequency categories into an 'Other' category.

Location	Country
LC-39A, Kennedy Space Center, Florida, USA	USA

Datum	Extracted_Year	Extracted_Month	Extracted_Time
Thu Apr 18, 1968 10:30 UTC	1968	4	10.50

The grouping step was crucial for stabilizing the model and mitigating the risk of overfitting on rare data points.

Final Pipeline and Validation Setup

Numerical Scaling

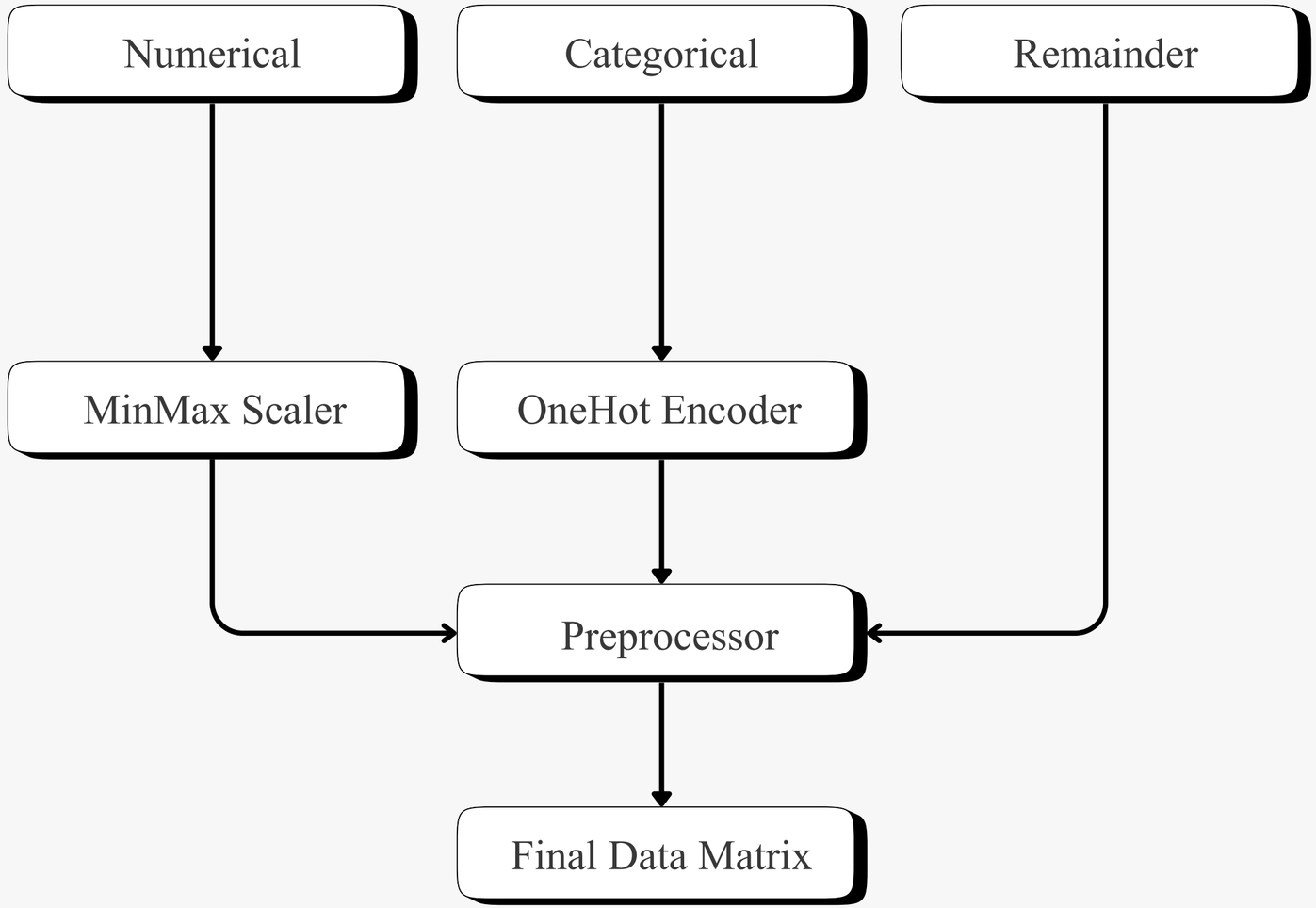
Applied MinMax Scaling to numerical features to normalize their values between 0 and 1

Categorical Encoding

Used One-Hot Encoding to convert categorical features into numerical vectors

Validation Strategy

Stratified Sampling for the Train/Validation split.



Baseline Model Evaluation

Approach:-Tested six baseline models to find the most suitable algorithm.

Initial Finding:- Tree-based models (XGBoost, LightGBM, Random Forest) showed superior performance, suggesting a non-linear relationship in the mission data.

Key Finding:-The XGBoost model provided the highest performance, making it the clear choice to dedicate further hyperparameter tuning efforts toward.

Model Algorithm	Validation F1-Score
XGBoost Classifier	**0.9523**
LightGBM Classifier	0.9495
Random Forest Classifier	0.9455
Logistic Regression	~0.9479
SVC (Support Vector)	~0.9479
KNN (K-Nearest Neighbors)	~0.9478

Hyperparameter Tuning & Model Selection

1

Baseline Testing

Six models evaluated

2

HPT on Top Models

(n_estimators=2000, eta=0.01, max_depth=6)

3

Ensemble Exploration

Bagging, Boosting, Voting methods tested

4

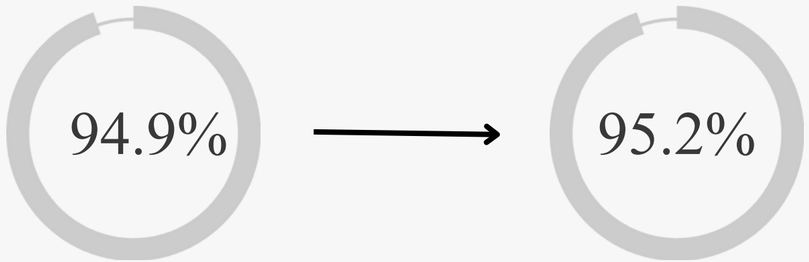
Champion Model

XGBoost selected

5

Cross Validation

Low model variance and good generalization



F1-Score

Deployment: Interactive Gradio Interface

Transformed the ML model into an accessible web application using Gradio, enabling instant mission outcome predictions.

User Input

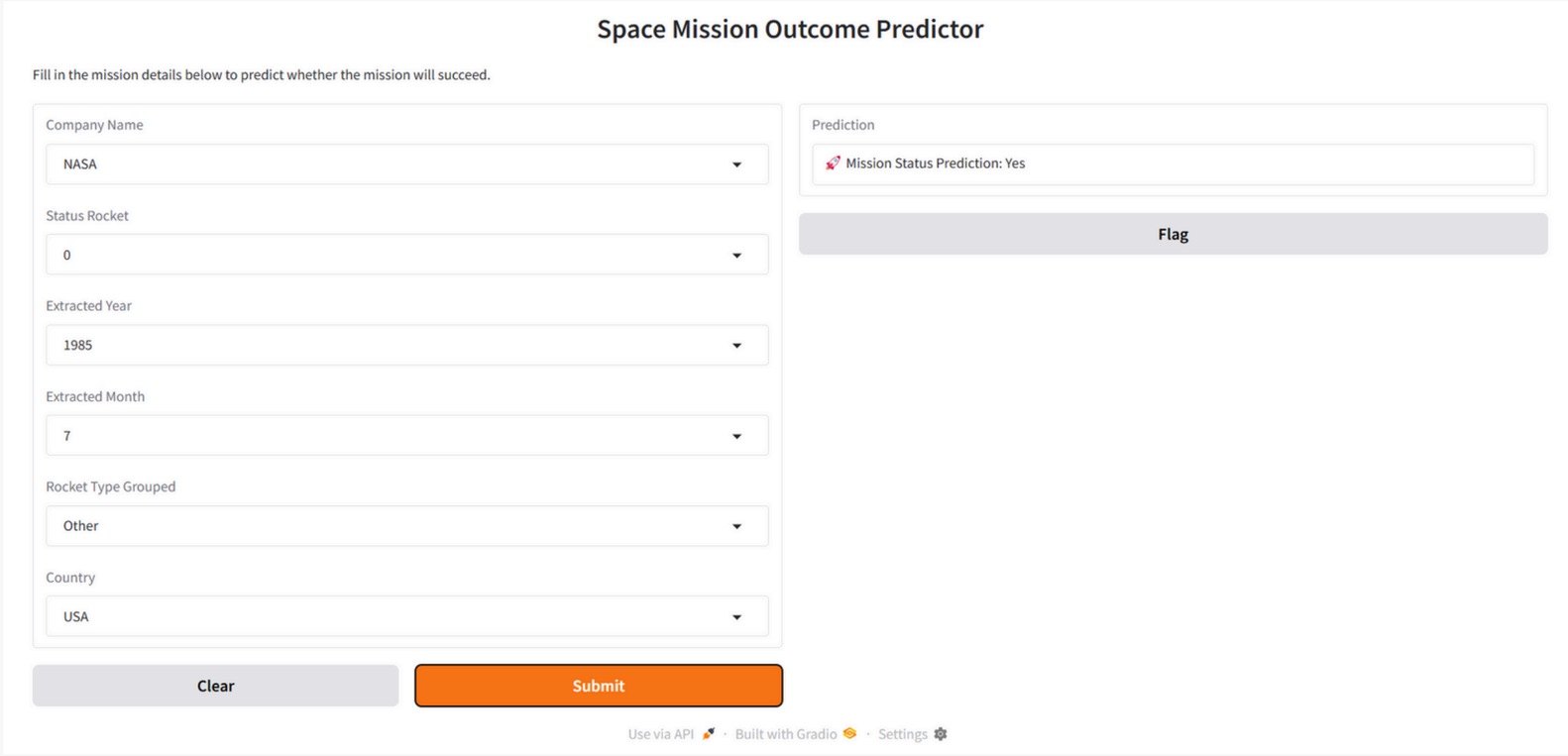
Select Company, Status, Year, Country from validated dropdowns

Backend Processing

Utilizes saved model2.pkl and pipeline.pkl for preprocessing

Instant Prediction

Returns mission success probability in real-time



Limitations & Future Scope

Project Summary

SkyCast demonstrates a comprehensive end-to-end machine learning solution with robust data cleaning, advanced feature engineering, and rigorous model selection achieving 95%+ F1-Score.

Limitations

Limited Metric Exploration

The project focused primarily on the F1-score and did not test other crucial metrics like ROC-AUC or Precision/Recall curves.

Explainability Gap

Lack of integrated SHAP/LIME tools prevents explaining why individual missions succeed or fail.

Future Enhancements

Model Explainability

Integrate SHAP (directly into the deployment application.

Deployment Monitoring

Implement continuous monitoring to detect data drift as new launch emerges.

Thank You!

Relevant links:-

1. SkyCast GitHub Repository Link :- https://github.com/Kitiksha1000/SkyCast
2. Kaggle Competition Link :- https://www.kaggle.com/competitions/sky-cast-margazhi-25
3. LinkedIn Profile Link :- https://www.linkedin.com/in/kitiksha-chakrawarti-7340ab289/