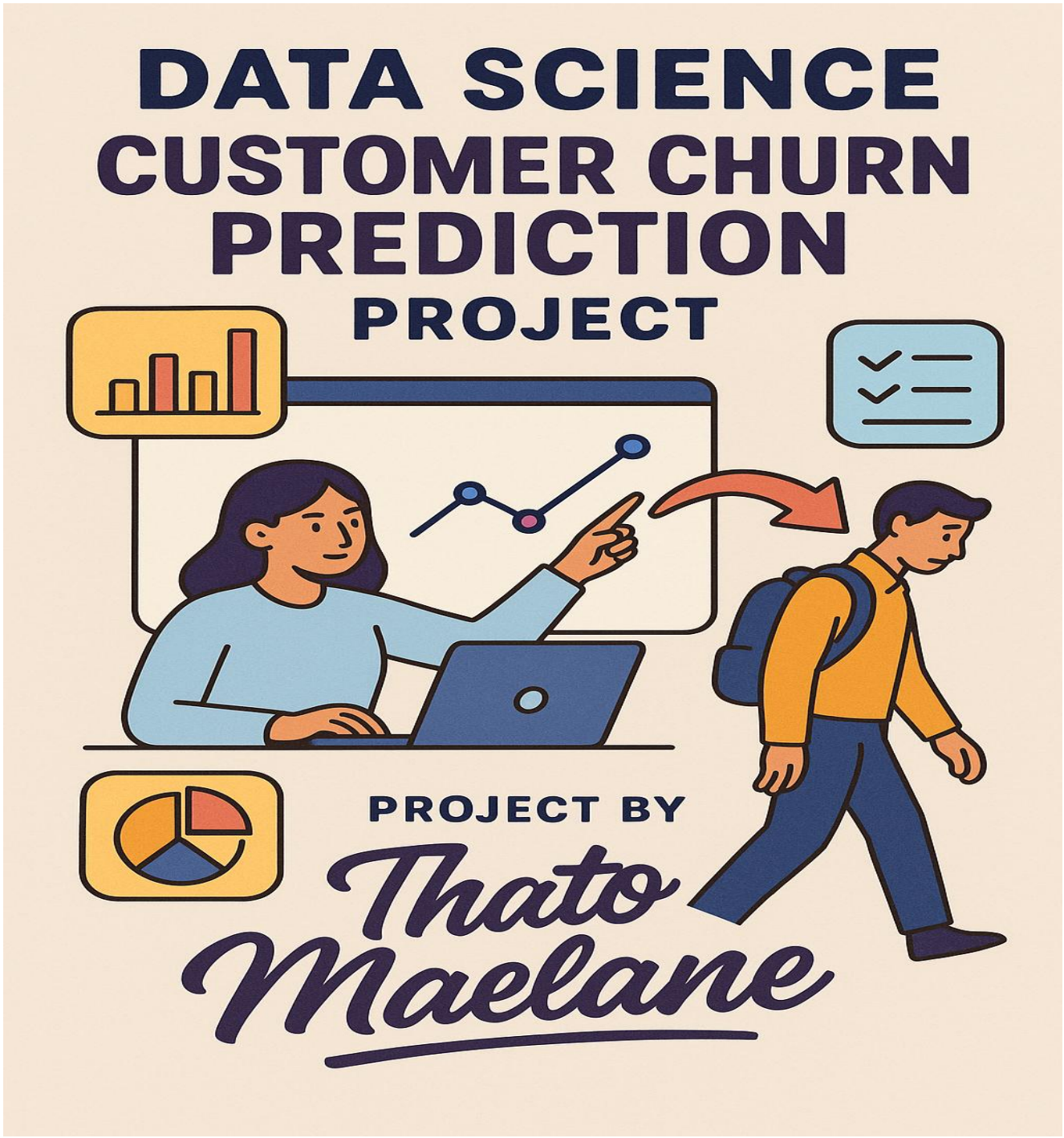


Name: Thato Maelane
Email: thato6216@gmail.com
LinkedIn: https://www.linkedin.com/in/thatomaelane
GitHub: https://github.com/thatomaelane
Kaggle: https://www.kaggle.com/thatomaelane
Date: 23 June 2025

Contents

Aim of the Project

The aim is to build a machine learning model that predicts whether a customer is likely to cancel (churn) their telecom service. This helps companies take proactive steps to retain customers and reduce revenue loss.

The Problem We're Solving

Customer churn directly impacts business profitability. By identifying customers who are likely to leave, businesses can take targeted actions like offering promotions or improving service. The challenge is to develop a model that can accurately classify churners based on their account and service usage data.

Dataset Information

Source: Public Telco Customer Churn dataset

Features: Customer demographics, services (like Internet, phone), account information (contract type, tenure), and charges.

Label: Churn column (Yes or No)

Preprocessing:

- Converted TotalCharges to numeric
- Removed irrelevant features like customerID
- Encoded categorical variables with pd.get_dummies()
- Standardized features using StandardScaler

Language Used

- Python

Models Used

1. Logistic Regression
2. Random Forest Classifier
3. XGBoost Classifier

Tools & Libraries

- **Data Processing**: pandas, numpy
- **Visualization**: seaborn, matplotlib
- **Modeling**: scikit-learn, XGBoost
- **Evaluation**: Confusion Matrix, Precision, Recall, F1-score

Results & Findings

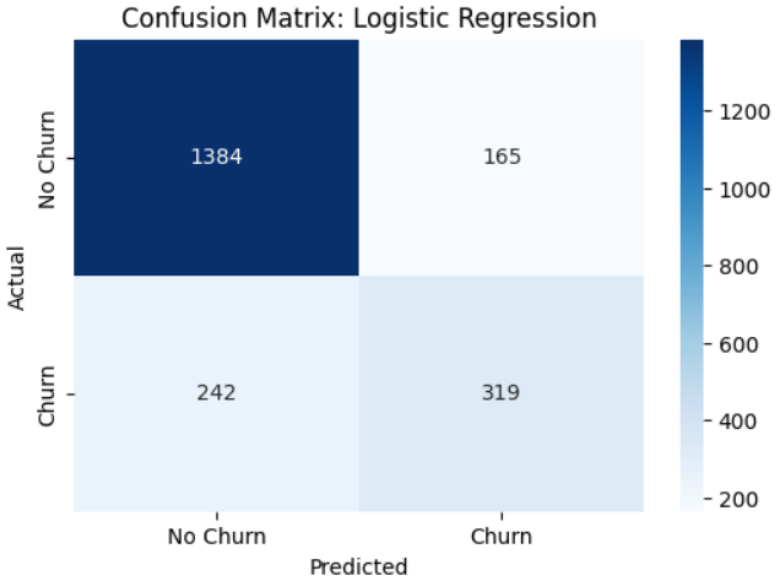
1. Logistic Regression

```
Confusion Matrix:
[[1384  165]
 [ 242  319]]
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.89      0.87      1549
           1       0.66      0.57      0.61       561

    accuracy                           0.81      2110
   macro avg       0.76      0.73      0.74      2110
weighted avg       0.80      0.81      0.80      2110
```



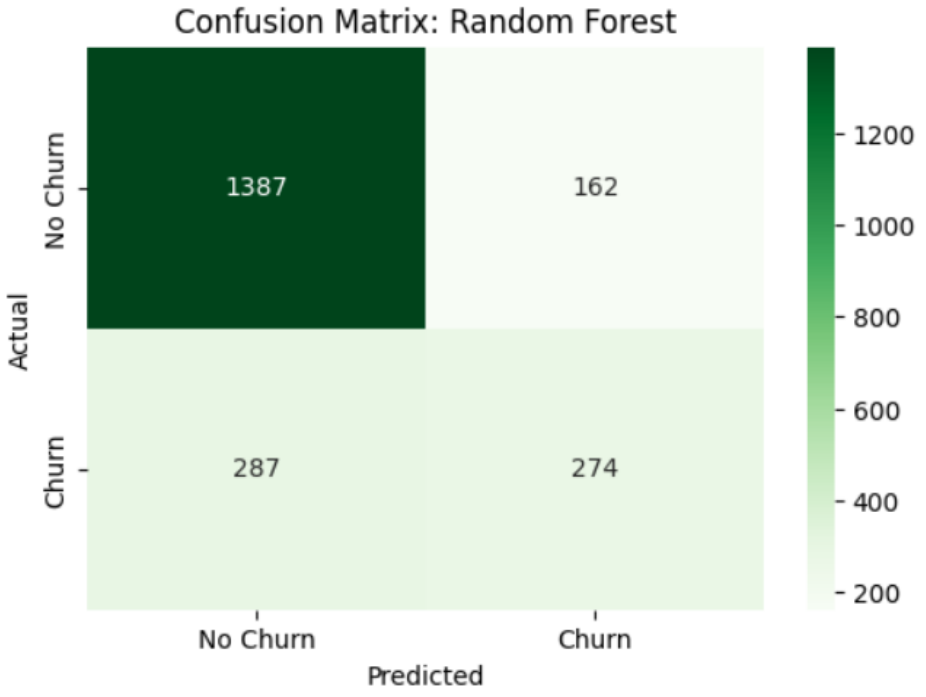
2. Random Forest

```
Random Forest Confusion Matrix:
[[1387  162]
 [ 287  274]]

 Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.90      0.86      1549
           1       0.63      0.49      0.55       561

    accuracy                           0.79      2110
   macro avg       0.73      0.69      0.71      2110
weighted avg       0.78      0.79      0.78      2110
```

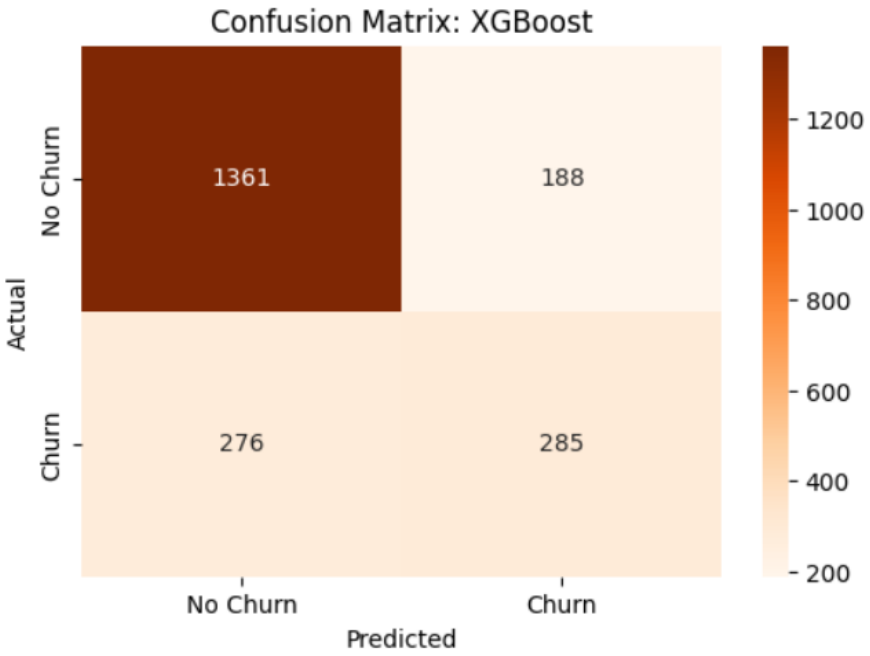


3. XGBoost

```
XGBoost Confusion Matrix:
[[1361  188]
 [ 276  285]]
```

XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.83	0.88	0.85	1549
1	0.60	0.51	0.55	561
accuracy			0.78	2110
macro avg	0.72	0.69	0.70	2110
weighted avg	0.77	0.78	0.77	2110



Model	Accuracy	Precision (Churn)	Recall (Churn)	F1-Score (Churn)
Logistic Regression	81%	0.66	0.57	0.61
Random Forest	79%	0.63	0.49	0.55
XGBoost	78%	0.60	0.51	0.55

Key Finding: Logistic Regression performed best overall for this dataset, offering a good balance between simplicity and effectiveness.

Summary & Conclusion

This project showed how machine learning can be applied to a real-world business problem predicting customer churn. Through data cleaning, feature engineering, model training, and evaluation, we developed a predictive system that gives telecom companies the insight needed to keep customers from leaving. Logistic Regression outperformed other models, demonstrating that sometimes simpler models offer better generalization for structured data.

References

- Dataset: Telco Customer Churn on Kaggle
- Scikit-learn documentation: https://scikit-learn.org
- XGBoost documentation: https://xgboost.readthedocs.io