Purwadhika's DTI-DS Program

Muhamad Farikhin

2024

Capstone Modules 3 & 4 (Machine Learning & Cloud Computing)

ECOMMERCE CUSTOMER CHURN

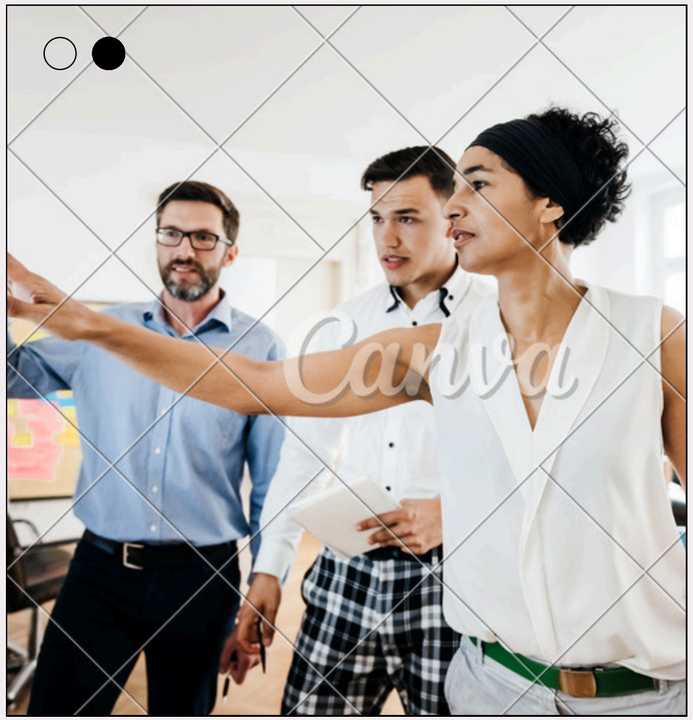
Using Machine Learning to Predict Customer Attrition for Better Decision Making at an E-commerce Company

Classification

Cloud Deployment

Optimal Strategy

Minimize costs while retaining the most customers



Background

Southeast Asia's internet economy is expected to grow from **$194 billion to over $330 billion by 2025**, with Indonesia leading at $82 billion in 2023 (US International Trade Administration). As online shopping becomes more common, consumers will likely expand their **digital purchases beyond apparel and low-value electronics** to food and beverage, beauty products, and home goods (McKinsey & Company). **Indonesia's e-commerce market is highly saturated and competitive**, facing threats from international players looking to expand their business in the country.

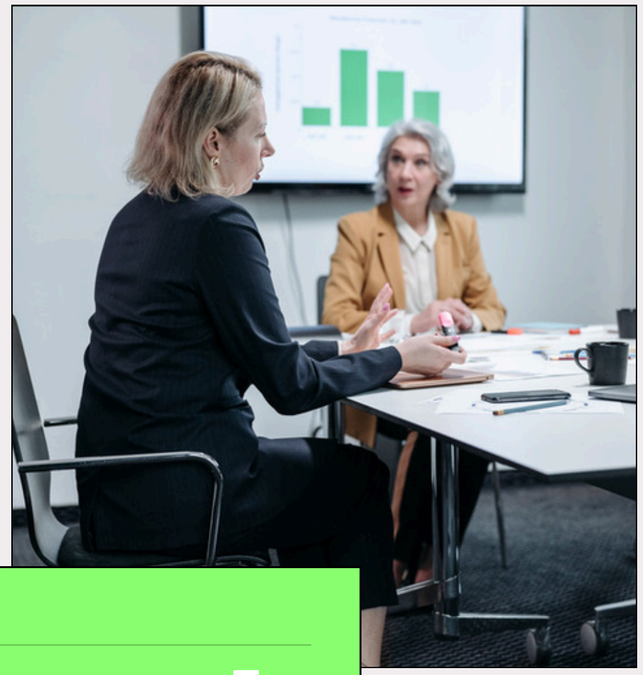
Implication:

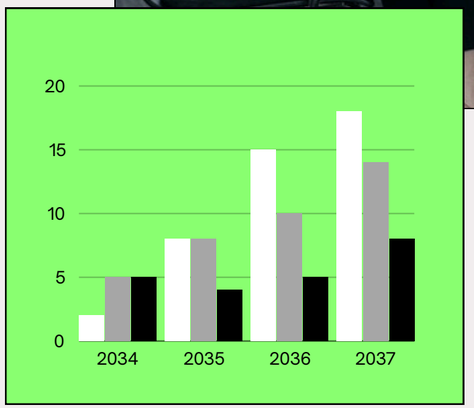
- Customers often switch between e-commerce platforms.
- Acquiring new customers is more costly than retaining existing ones.

Solution:

- Identify potential churners and implement effective retention strategies.
- Understand the reasons behind customer churn to develop targeted win-back tactics.

Bain & Company: "A mere 5% increase in customer retention can boost profits by 25% to 95%"





Analysis Summary

- Churning customers have a median tenure of 3 years, compared to 9 years for non-churning ones, with churn rates decreasing over a decade. **Among customers with 5 years or less tenure, churning customers are half the number of non-churning ones.**
- **Complained customers exhibit higher churn rates** than those who haven't complained.
- **Customers receiving lower cashback amounts have higher churn rates** compared to those receiving higher cashback amounts.

Data Understanding

The dataset consists of **user information**,

Attributes	Description
Tenure	Tenure of a customer in the company.
WarehouseToHome	Distance between the warehouse to the customer's home.
NumberOfDeviceRegistered	Total number of deceives is registered on a particular customer.
MaritalStatus	Marital status of a customer.
NumberOfAddress	Total number of address on a particular customer.

and **transaction-related data**

Attributes	Description
PreferedOrderCat	Preferred order category of a customer in the last month.
SatisfactionScore	Satisfactory score of a customer on service.
Complain	Any complaint has been raised in the last month.
DaySinceLastOrder	Day since last order by customer.
CashbackAmount	Discount offered on that specific product.

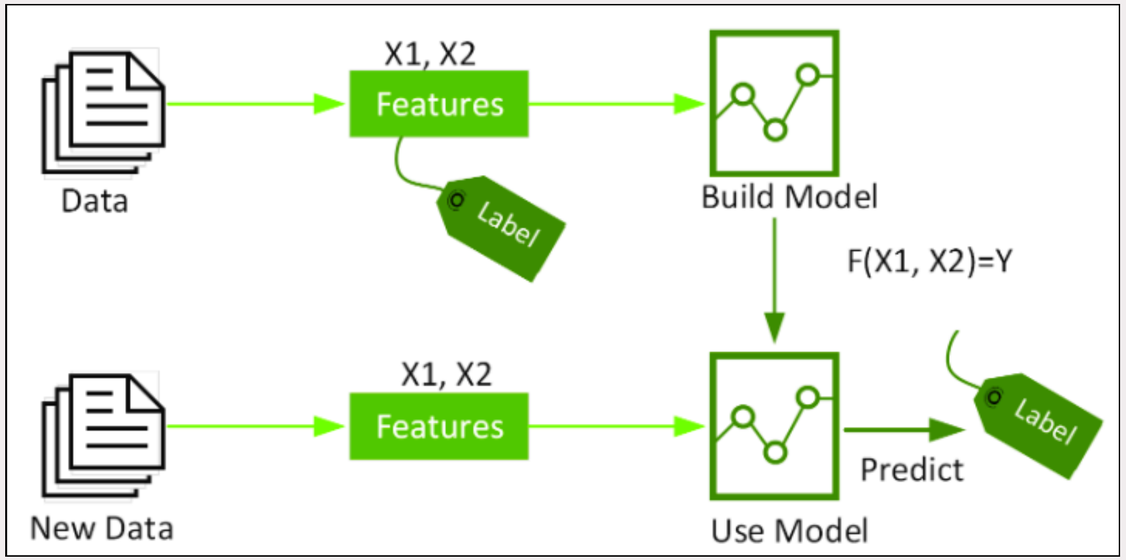
Each record represents a user profile and their transactions, indicating whether the user has churned.

Machine Learning Implementation

1. Leveraging a Powerhouse Technology
Machine learning can uncover hidden patterns and process vast amounts of data, providing significant benefits for companies that use it effectively.

2. Investment Worth Making
While integrating machine learning into decision-making involves upfront costs, when done correctly, it will help prevent customer loss and navigate the creation of effective retention strategies.

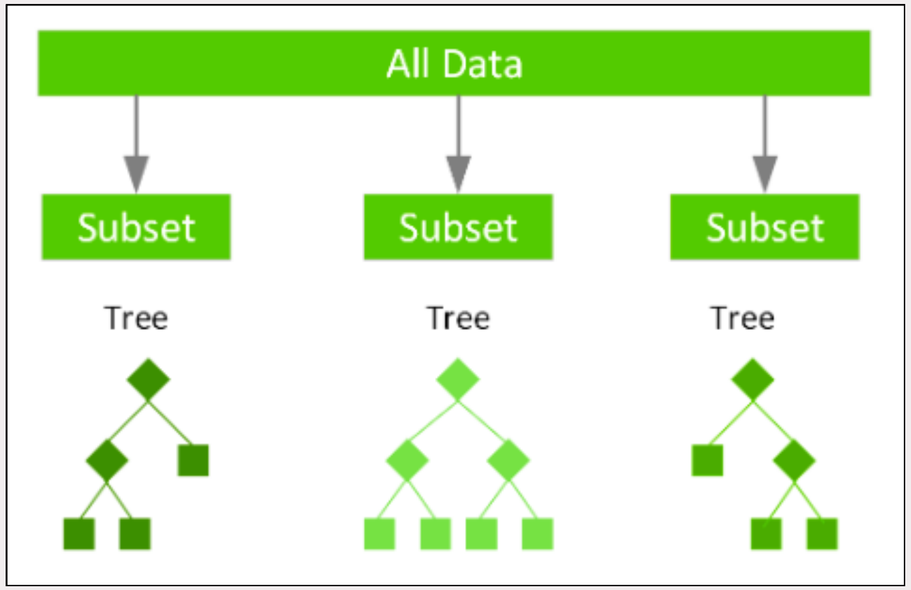
3. Churn Classification in Industry
Algorithms like XGBoost are advanced tools for churn prediction and claim processing in insurance. They provide accurate results by capturing complex, non-linear relationships in the data.



XGBoost

Powerful machine learning algorithm offering versatility and performance:

- **Boosting Technique:** Builds a sequence of decision trees, where each tree corrects errors from the previous ones, enhancing prediction accuracy.
- **Gradient Descent:** Adjusts tree weights iteratively to minimize model error.
- **Regularization:** Uses L1 and L2 techniques to prevent overfitting and improve generalization.
- **Parallel Processing:** Speeds up computations, making it efficient for large datasets.



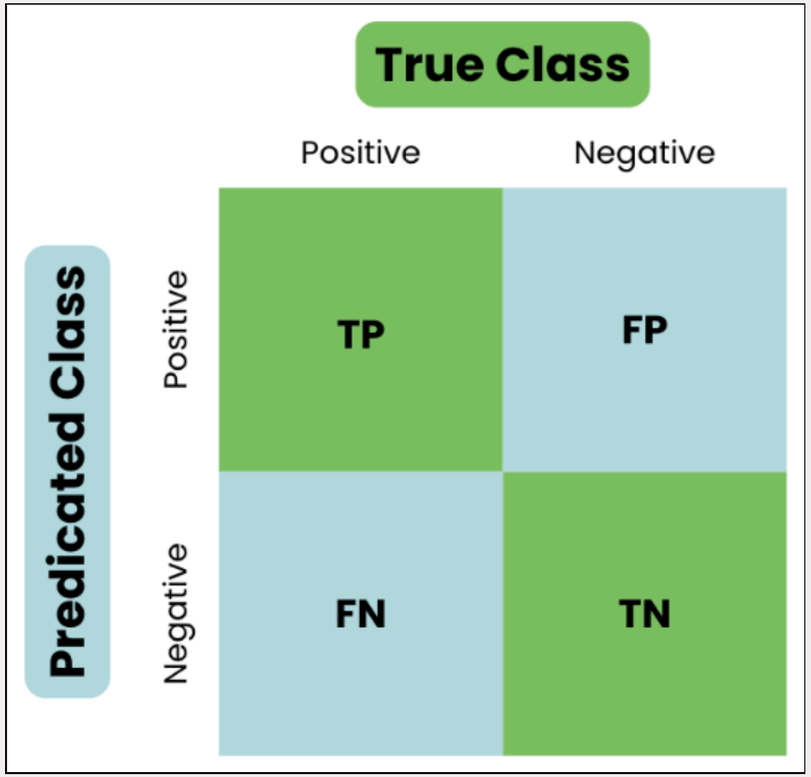
Performance Metric

The F2 score is a weighted harmonic mean of Precision and Recall, giving twice as much weight to Recall compared to Precision. This aims to:

$$\text{F2 Score} = (1 + 2^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(2^2 \cdot \text{Precision}) + \text{Recall}}$$

This approach is suited for this problem because it:

- **Minimize False Negatives**: The F2 score emphasizes reducing the number of customers misidentified as not churning.
- **High Recall Focus**: A high F2 score indicates that the model effectively identifies most customers who are churning, even if it occasionally misclassifies some non-churning customers as churning.



Data Preprocessing

Data Clearning Steps:

Process	Description
Drop duplicate values	Remove any duplicate entries to ensure data integrity.
Check nullity & unique values	Check for nullity and unique values, apply descriptive analysis, and examine the values of categorical attributes.
Handle missing values	Fill in missing values using the median for skewed distributions.
Handle outliers	Use Interquartile Range (IQR) to detect and remove only contextual outliers for '*Tenure*', '*WarehouseToHome*', '*NumberOfAddress*'.

Feature Engineering & Selection Steps:

Process	Description
Encoding	Apply Binary Encoding for '*PreferredOrderCat*' and '*Marital Status*'
Normalization	Apply MinMaxScaler (0 to 1) to preserve the shape of the original distribution while bringing values within a desired range, especially for non-normal distributions
Recursive Feature Elimination (RFE)	Apply RFE to select important features after identifying the best model.

Models Performance Comparison

Model for Benchmarking:

Base Models:

- Logistic Regression
- K-Nearest Neighbors
- Decision Tree

Ensemble Models:

- Various Types:
  - Voting (Base Models)
  - Stacking (Base Models)
- Same Type (Bagging):
  - Random Forest
- Same Type (Boosting):
  - AdaBoost
  - Gradient Boosting
  - XGBoost

Model	Train Score	Test Score	Difference
XGBoost	**0.63**	**0.70**	0.06
Random Forest	0.56	0.65	0.08
ADABoost	0.58	0.65	0.06
Gradient Boosting	0.56	0.65	0.08
Stacking	0.54	0.63	0.08
Decision Tree	0.62	0.62	**0.003**
Voting	0.53	0.60	0.06
Logistice Regression	0.44	0.49	0.04
KNearestNeighbor	0.43	0.48	0.05

Model Evaluation

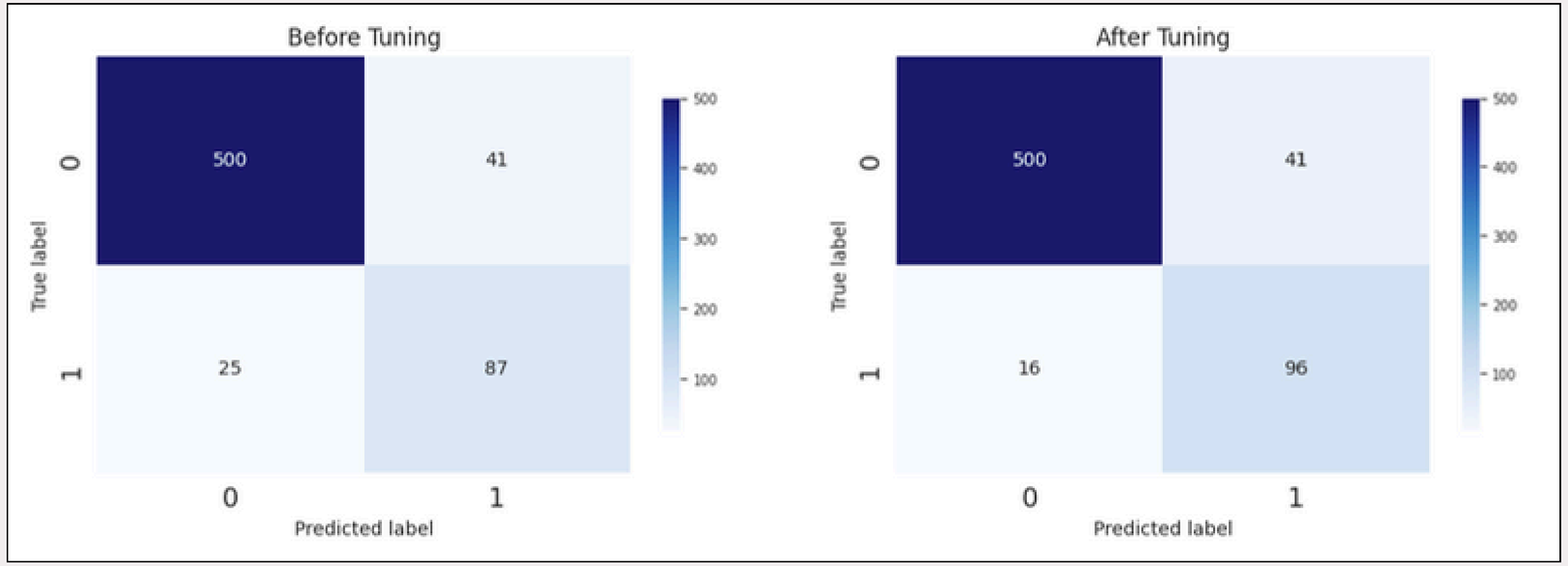
Model Pipeline:

1. **Encoding:** Handle categorical data through encoding.
2. **Imbalance Treatments:**
   - Cost-Sensitive Learning:
     - XGBoost: Set scale_pos_weight to 5.2.
     - Decision Tree: Use class_weight with {0: 1, 1: 5}.
   - Resampling: Apply Random Under Sampling for both XGBoost and Decision Tree.
3. **Hyperparameter Tuning:** Perform tuning for both models.

Model	F2-Score
XGBoost with Random Under Sampling	0.77
Weighted XGBoost	**0.82**
Decision Tree with Random Under Sampling	0.76
Weighted Decision Tree	0.68

Model Evaluation - Best Model After Tuning

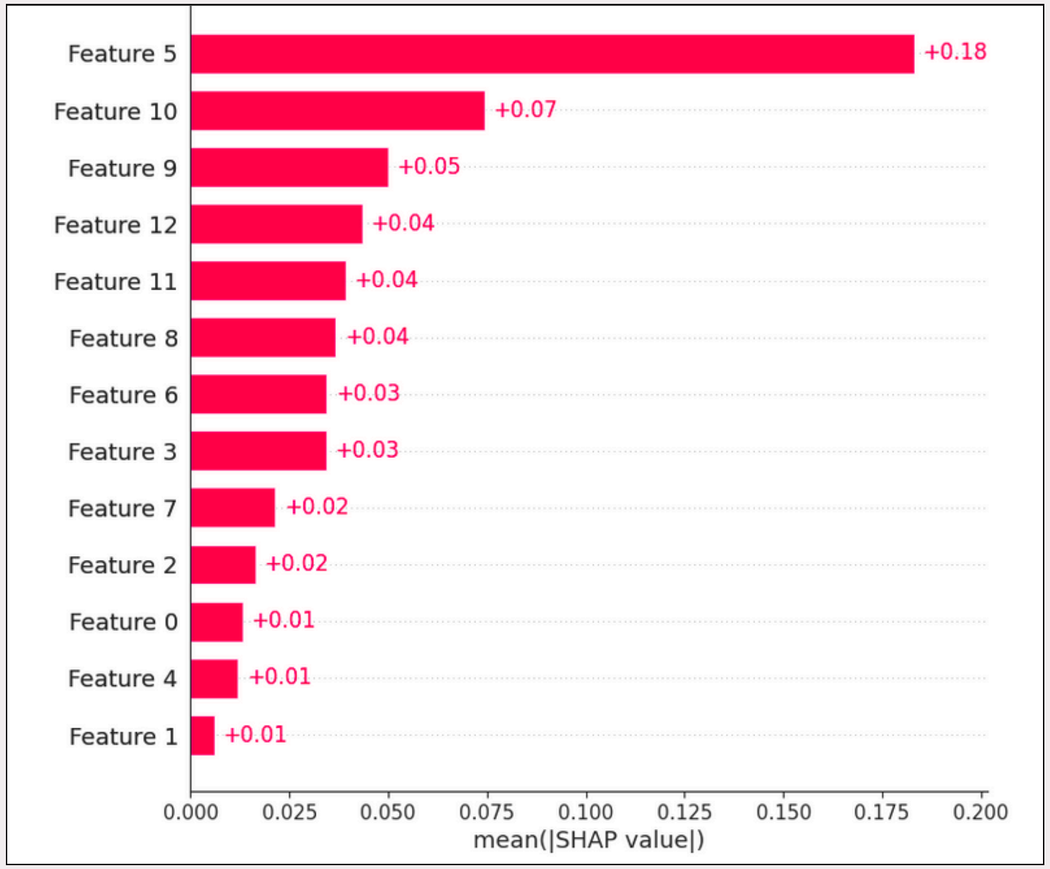
Parameter	Value
booster	'dart'
colsample_bytree	0.6956
gamma	1.7302
learning_rate	0.1627
max_depth	9
min_child_weight	2
n_estimators	86
reg_alpha	1.8586
reg_lambda	0.6371
subsample	0.8337



The number of **False Negatives** decreases significantly after tuning.

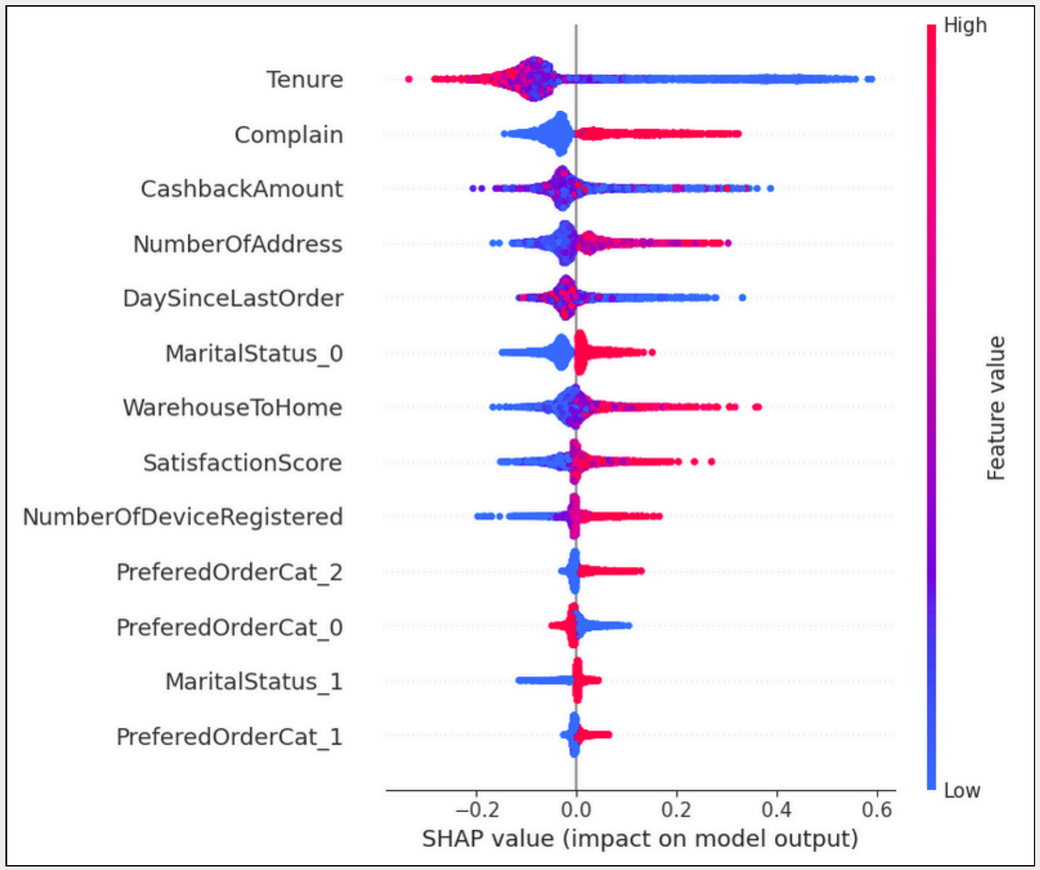
Model Interpretation

SHAP Bar Plot



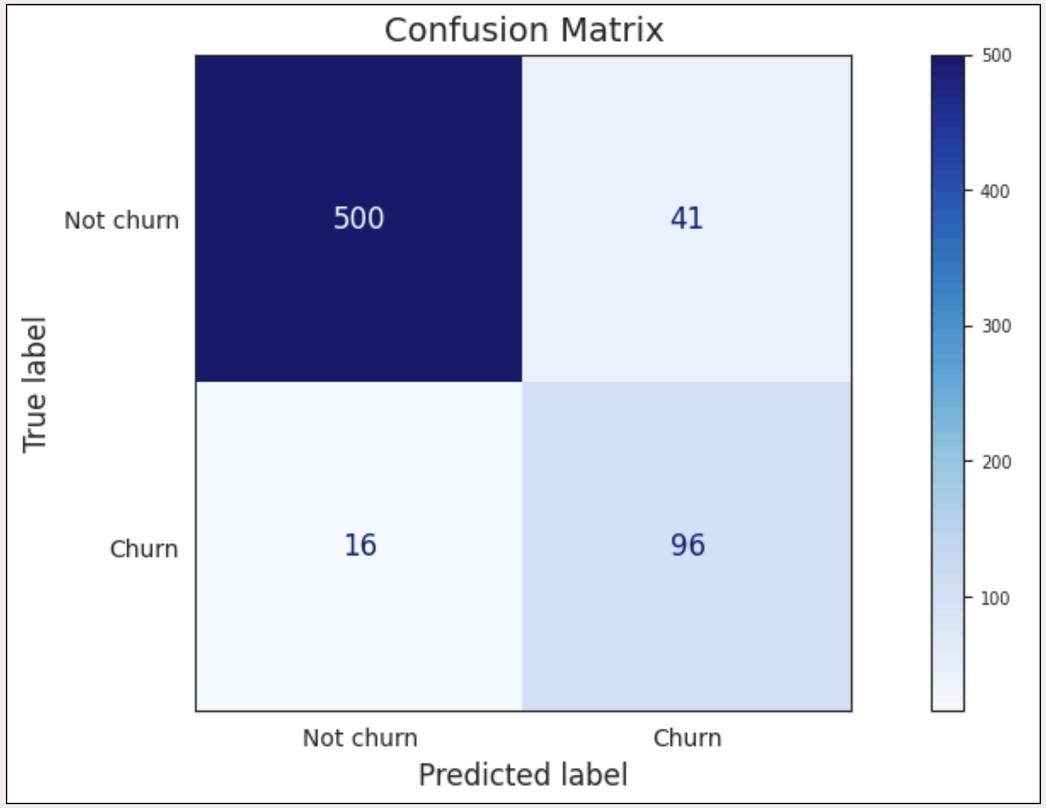
- **Tenure** (Feature 5): Most influential, contributing ~0.18 per prediction.
- **Complain** (Feature 10): Contributes ~0.07.
- **Cashback Amount** (Feature 9): Contributes ~0.05.

SHAP Summary Plot



- **Tenure**: Low tenure customers are more likely to churn.
- **Complaints**: Customers who submit complaints are more likely to churn.
- **Cashback Amount**: Saturated values significantly influence predictions, indicating distinct user responses to cashback levels.

Cost Benefit Analysis



Assumptions:

CAC (Customer Acquisition Cost) = USD 75
CRC (Customer Retention Cost) = USD 15

Churn Loss Without ML Model:

Churn Loss = Churn Rate x CAC
Churn Rate = 16%
Churn Loss = 0.16 x 75
Churn Loss = USD 8,025

Churn Loss With ML Model:

Churn Loss = (FN x CAC) + ((TP + FP) x CRC)
FN (False Negatives) = 16
TP (True Positives) = 96
FP (False Positives) = 41
Churn Loss = (16 x 75) + ((96 + 41) x 15)
Churn Loss = USD 3,255

Potential Savings:

Amount: **USD 4,770**
Rate: **59.44%**

Conclusion

This model can guide strategic decisions, improve customer retention, and enhance overall business performance by first predicting which customers are likely to churn and then intervening before they leave. Some possible recommendations for intervention and maximizing retention include:

1. To improve perceived value and maximize retention, offer loyalty promotion programs to customers predicted to churn by the model, especially those with a tenure of five years or less.
2. Prioritize resolving customer complaints, as analysis shows that those likely to churn often have unresolved issues.
3. The analysis shows that churning customers receive lower average cashback values compared to non-churning customers. The company should consider increasing cashback amounts for those at risk of churning.

Running the Model on the Cloud

Steps:

1. Set up a Python virtual environment with the necessary libraries for model prediction.
2. Prepare the data for bulk prediction.
3. Run the prediction.
4. Save the prediction results to BigQuery.

Thank You

"Errare humanum est" — "To err is human."