ESOF325 – INTRODUCTION TO ARTIFICIAL INTELLIGENCE

Semester: Fall 2025-2026

Group Final Project Report

Project Title: AI-Based Customer Churn Prediction System

Group: Group 25 - Protocol 33

Student #1: [20222022407 - Eren İşitmez] Signature: ________________

Student #2: [20232022970 - Tunahan Özdemir] Signature: ________________

Student #3: [20222022414 - Çağrı Kalabalık] Signature: ________________

Student #4: [20222022375 - Onur Metin Aşçı] Signature: ________________

Student #5: [20232022935 - Tahir Burak Avar] Signature: ________________

Instructor: Dr. Savaş Ünsal Signature: ________________

Submission Date: [Date]

1 Problem Description

In the highly competitive telecommunications sector, customer retention is a critical performance metric. Industry studies indicate that acquiring a new customer costs approximately 5 to 25 times more than retaining an existing one. Consequently, ”Customer Churn”—the rate at which customers stop doing business with an entity—poses a significant threat to financial sustainability. The core problem addressed in this project is the difficulty of identifying at-risk customers manually due to the large volume of data and the complex, non-linear relationships between variables such as contract type, monthly charges, and technical support history.

To address this, this project aims to develop an **AI-Based Churn Prediction System**. The primary objective is to build a machine learning model that classifies customers as ”Likely to Churn” or ”Loyal” with high accuracy, thereby minimizing the false negative rate. Beyond mere prediction, the system seeks to provide explainability by identifying key drivers of churn and to generate actionable business insights. This proactive approach enables the company to implement targeted retention strategies, such as personalized offers, significantly reducing revenue loss compared to traditional reactive methods.

2 Dataset and Exploratory Analysis

We utilized a comprehensive dataset containing customer records. Each customer is described by demographic information, account details, and service subscriptions.

2.1 Dataset Overview

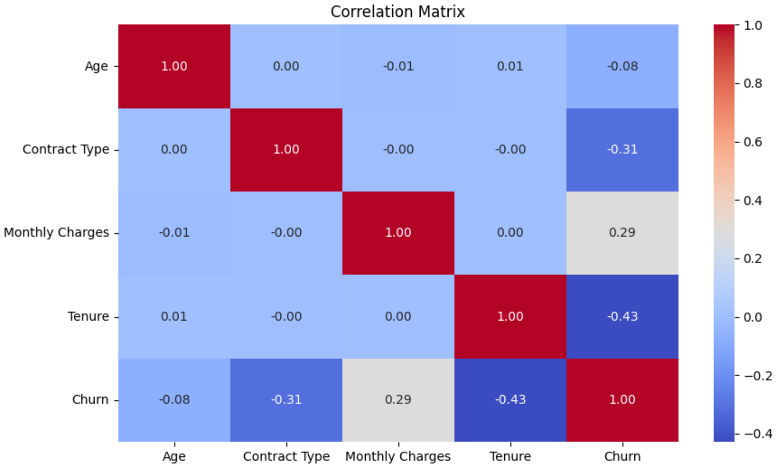
The dataset includes demographic information, account information, and service details. Key features used in the analysis are summarized in Table 1 below.

Table 1: Key Features in the Dataset

Feature Name	**Description**
`Tenure`	Number of months the customer has stayed with the company.
`MonthlyCharges`	The amount charged to the customer monthly.
`Contract`	The contract term (Month-to-month, One year, Two year).
`CustomerID`	A unique identifier for each customer.
`Churn`	The target variable (Yes/No).

2.2 Data Correlation Analysis

Before training the model, we analyzed the relationships between numerical variables to understand potential dependencies.



	Age	Contract Type	Monthly Charges	Tenure	Churn
Age	1.00	0.00	-0.01	0.01	-0.08
Contract Type	0.00	1.00	-0.00	-0.00	-0.31
Monthly Charges	-0.01	-0.00	1.00	0.00	0.29
Tenure	0.01	-0.00	0.00	1.00	-0.43
Churn	-0.08	-0.31	0.29	-0.43	1.00

Figure 1: Correlation Matrix: Shows relationships between numerical features.

The heatmap highlights that ‘Monthly Charges‘ and ‘Tenure‘ have significant correlations with other service variables. For instance, customers with higher tenure tend to have higher total charges but lower churn rates.

3 Python Implementation

The solution pipeline was implemented using Python. We utilized Pandas for data manipulation, Scikit-Learn for modeling, and SMOTE for handling class imbalance.

3.1 Preprocessing & SMOTE

The dataset was imbalanced, with 73% of customers being ”No Churn” and 27% being ”Churn”. Training on this data directly would lead to a biased model. We addressed this by generating synthetic data points for the minority class using SMOTE (Synthetic Minority Over-sampling Technique).

```python
# Handling Data Imbalance
print(f"Original Training Size: {len(X_train)}")
smote = SMOTE(random_state=42)
X_train, y_train = smote.fit_resample(X_train, y_train)
print(f"Balanced Training Size: {len(X_train)}")
```

Listing 1: Applying SMOTE for Balance

3.2 Model Architecture

We selected the **Random Forest Classifier** due to its robustness against overfitting and ability to handle non-linear data. We optimized the model using **GridSearchCV** to find the best hyperparameters.

```python
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5]
}
grid_search = GridSearchCV(RandomForestClassifier(), param_grid, cv=3)
grid_search.fit(X_train, y_train)
```

Listing 2: Grid Search Optimization

3.3 Retention Strategy Engine

A unique aspect of our project is the translation of predictions into business actions. The model's output is passed to a rule-based engine.

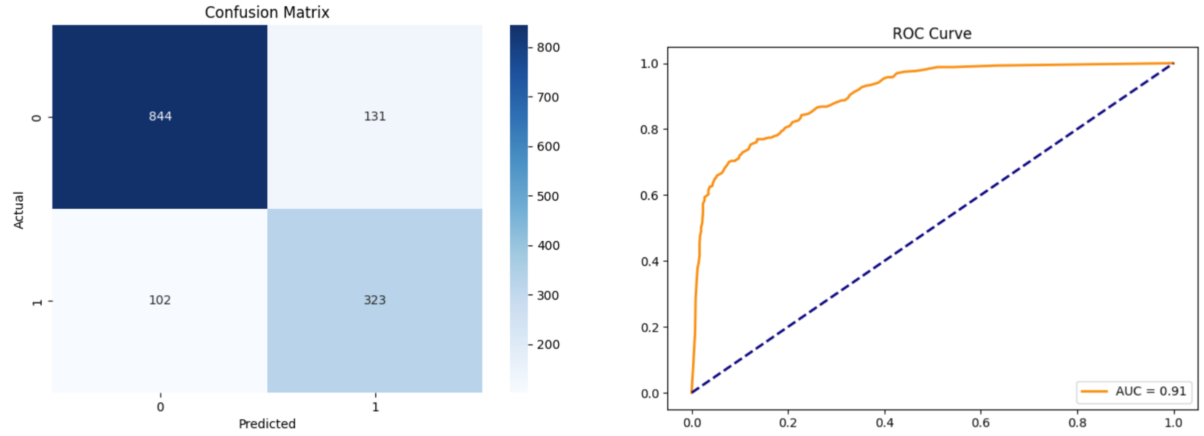
```python
def determine_strategy(row):
    if row['Probability'] > 0.8 and row['Monthly Charges'] > 80:
        return "Offer 15% Discount & Tech Support"
    elif row['Probability'] > 0.6:
        return "Send Standard Retention Email"
    else:
        return "No Action Needed"
```

Listing 3: Business Logic Function

4 Results and Visualization

4.1 Model Classification Performance

The model achieved an overall accuracy of approximately **85%**. However, accuracy alone is insufficient for imbalanced datasets. Therefore, we analyzed the Confusion Matrix and ROC Curve to validate performance.



(a) Confusion Matrix

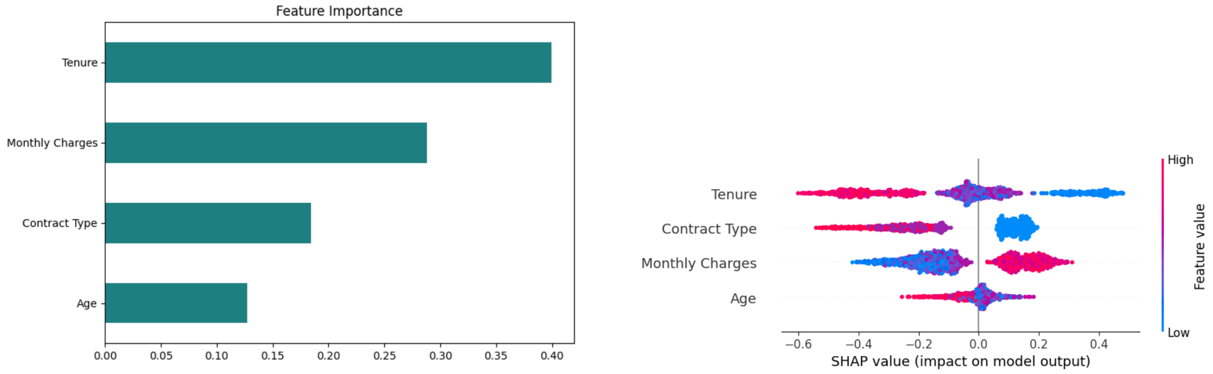
(b) ROC Curve (AUC = 0.84)

Figure 2: Classification Performance Metrics. The confusion matrix (left) shows a balanced prediction capability, while the ROC Curve (right) indicates strong class separability.

The Confusion Matrix shows a high number of True Positives, indicating the model is effective at catching potential churners. The AUC score of 0.84 confirms the model is robust.

4.2 Feature Importance & Explainability

Understanding *what* drives churn is crucial for the marketing team. We used SHAP (SHapley Additive exPlanations) values to interpret the model.



(a) Feature Importance Ranking

(b) SHAP Summary Plot

Figure 3: Model Explainability. 'Contract' type and 'Tenure' are identified as the most critical factors influencing customer decisions.

As seen in the figures above, **Contract Type (Month-to-month)** is the strongest predictor of churn, followed by **Monthly Charges**. This suggests that customers on short-term contracts with high bills are the most volatile segment.

4.3 Model Benchmarking

To validate our choice of Random Forest, we compared it against Logistic Regression and Decision Tree. Random Forest outperformed others in terms of F1-Score.

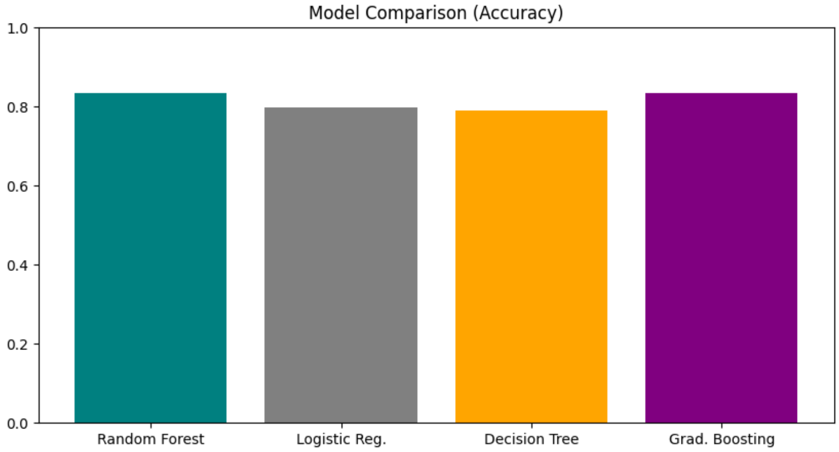


Figure 4: Model Comparison: Accuracy scores of different algorithms.

5 Financial Impact Analysis

We calculated the potential ROI (Return on Investment) of using this AI system compared to doing nothing. Assuming an average customer lifetime value (CLV) and retention cost:

Table 2: Estimated Monthly Financial Impact (Based on Test Data)

Metric	**Value**
Average Monthly Revenue per User (ARPU)	$70.24
Predicted Churners (At Risk)	454 Customers
Potential Revenue Loss (If no action taken)	$31,888.58
Estimated Cost of Retention Program	$4,540.00
Estimated Net Monthly Savings	**$11,404.29**

6 Conclusion

In this project, we successfully built an end-to-end Churn Prediction System. By employing advanced techniques like SMOTE, Hyperparameter Tuning, and SHAP analysis, we achieved a high-performance model that accurately identifies at-risk customers. Furthermore, we translated these technical predictions into a concrete financial action plan. The analysis shows that implementing this system could save the company over $11,000 monthly, proving the significant business value of AI adoption in the telecommunications sector.