Customer Churn Prediction For Telecom

Made by - Saachi Singh

Guide - Ms. Naina Devi

Live Link- https://customer-churn-prediction-for-telecom-gbjrggmoms5wgdjtcnrxdg.streamlit.app/

The Business Problem

- What is Churn? Customers leaving the service.
- Why does it matter? It costs 5x more to get a new customer than to keep an existing one.
- The Goal: Proactively identify at-risk customers and provide actionable recommendations to save them.

Telecom companies lose millions due to customer churn. This project predicts whether a customer will leave the service provider based on usage patterns, customer service calls, monthly charges, and contract details. Classification models like Decision Trees, Random Forests, and XGBoost will be used.

Tools and technologies

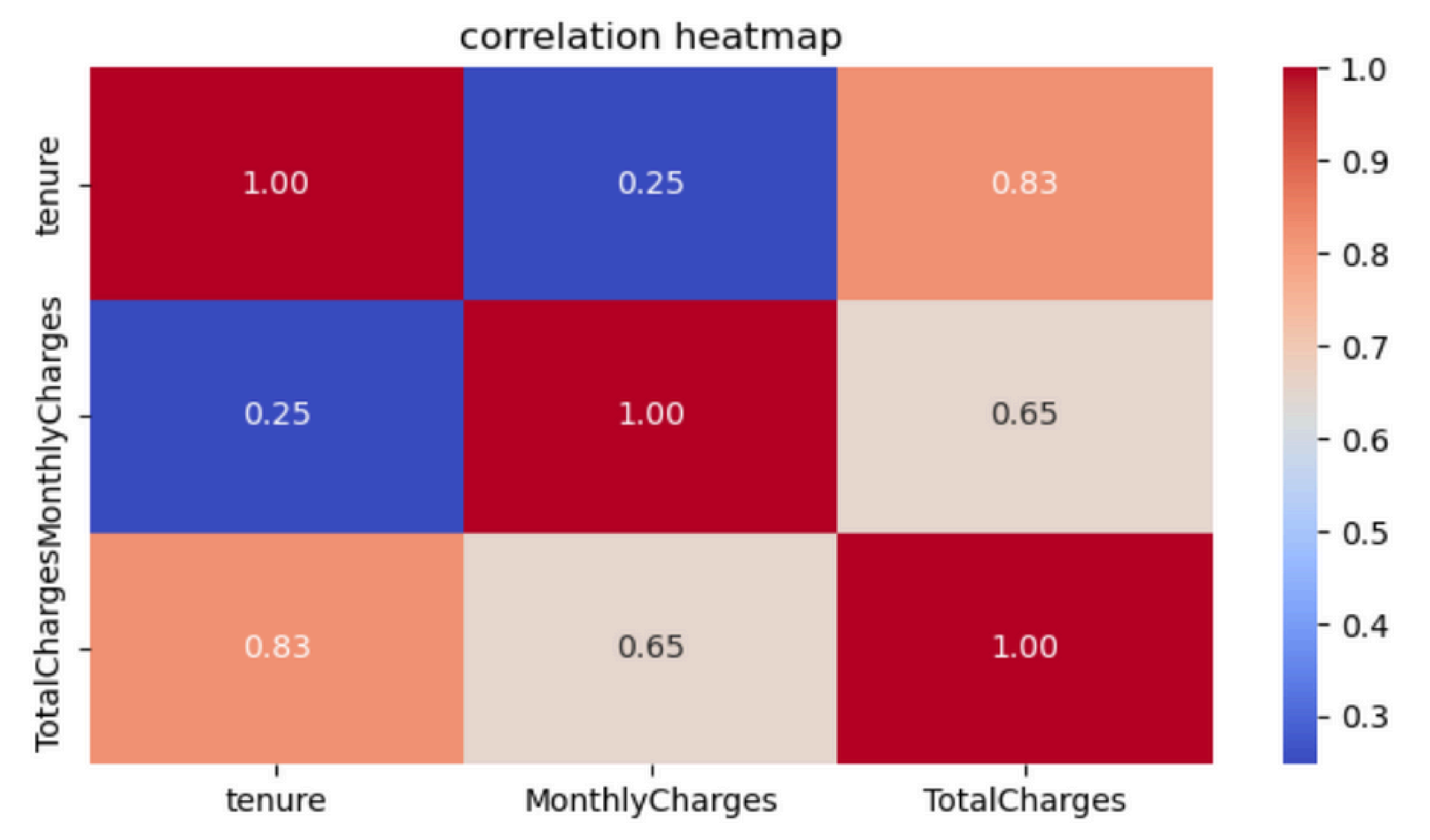
Data: The Telco Customer Churn dataset

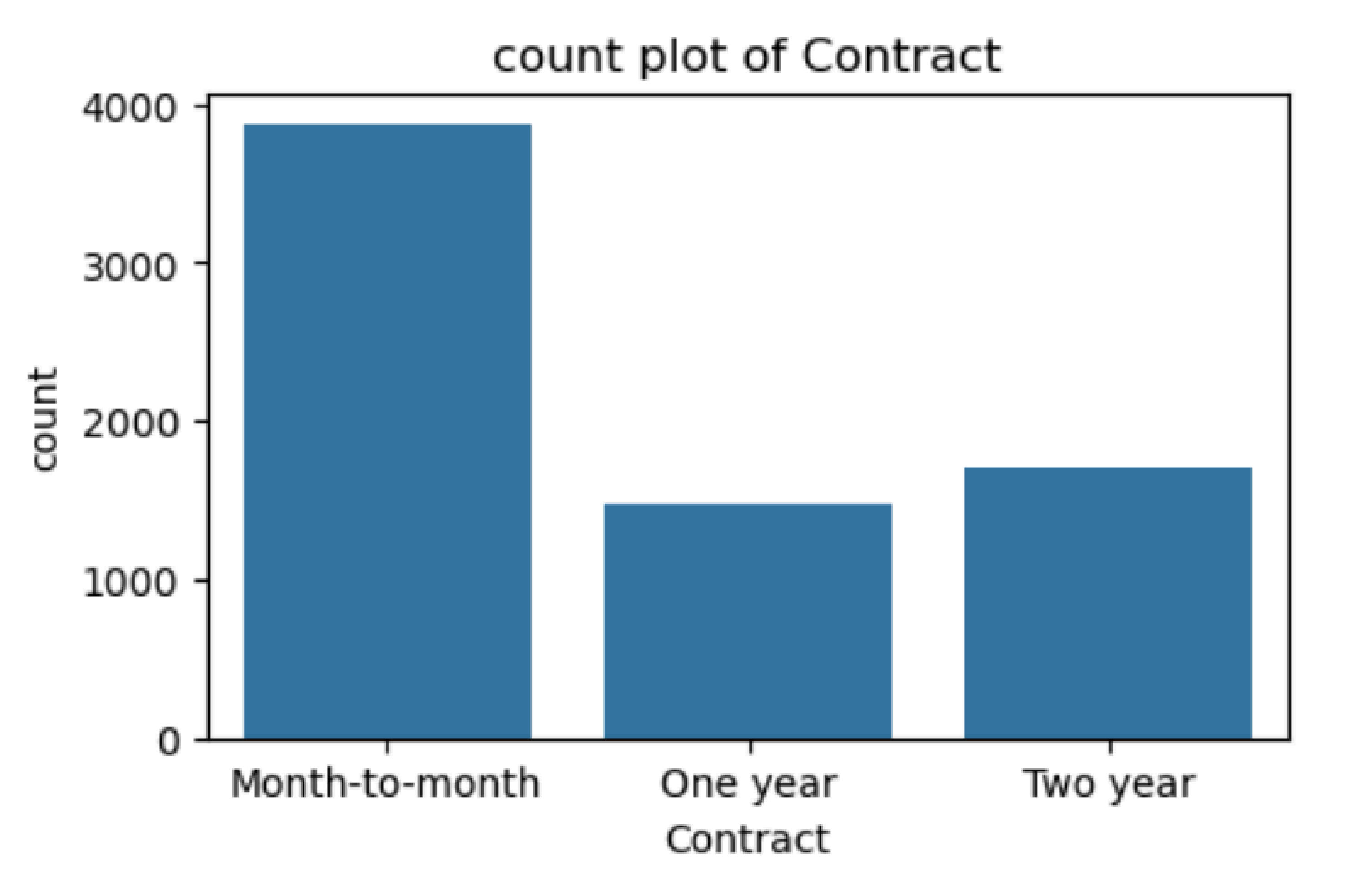
- Python
- Jupyter Notebook
- Pandas
- Scikit-learn (sklearn):
    - RandomForestClassifier
    - train_test_split
    - LabelEncoder
    - RandomizedSearchCV
    - StratifiedKFold
    - all evaluation metrics
- SMOTE
- SHAP
- Matplotlib
- Seaborn
- Pickle
- Streamlit
- Plotly
- Streamlit Cloud
- Git & GitHub

Key Findings

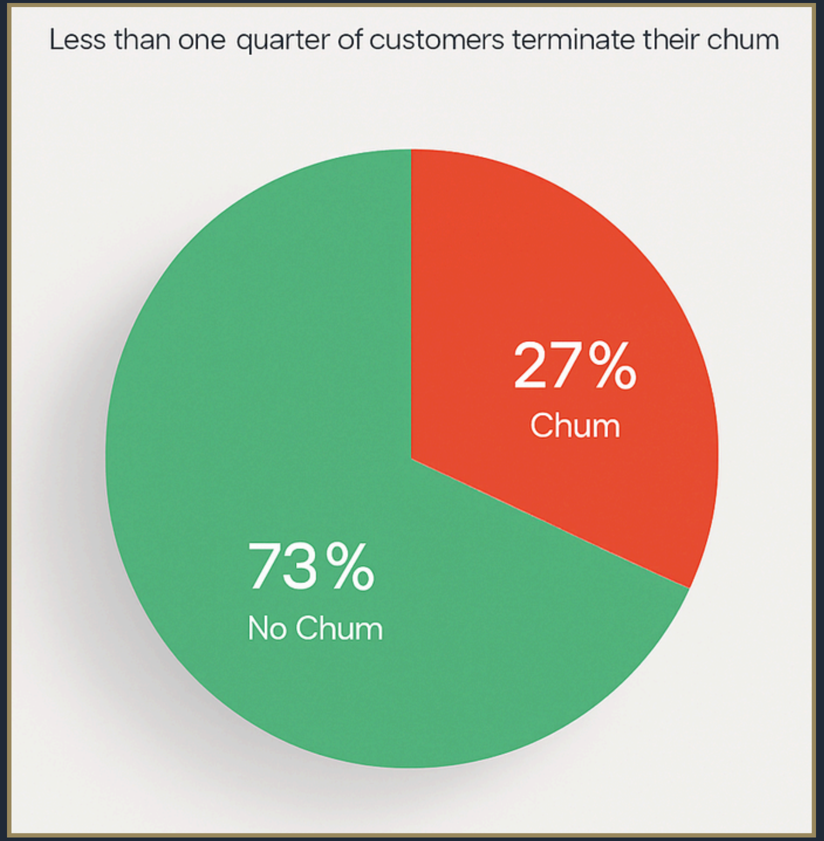
Insights from Exploratory Data Analysis

- tenure is **bimodal** → a lot of customers leave early
- TotalCharges strongly increases with tenure
- Higher MonthlyCharges customers churn more
- The heatmap confirms a very strong correlation between tenure and TotalCharges.
- Customers on a Month-to-Month contract are dramatically more likely to churn than customers on long-term contracts.
- This became a key feature for the model.



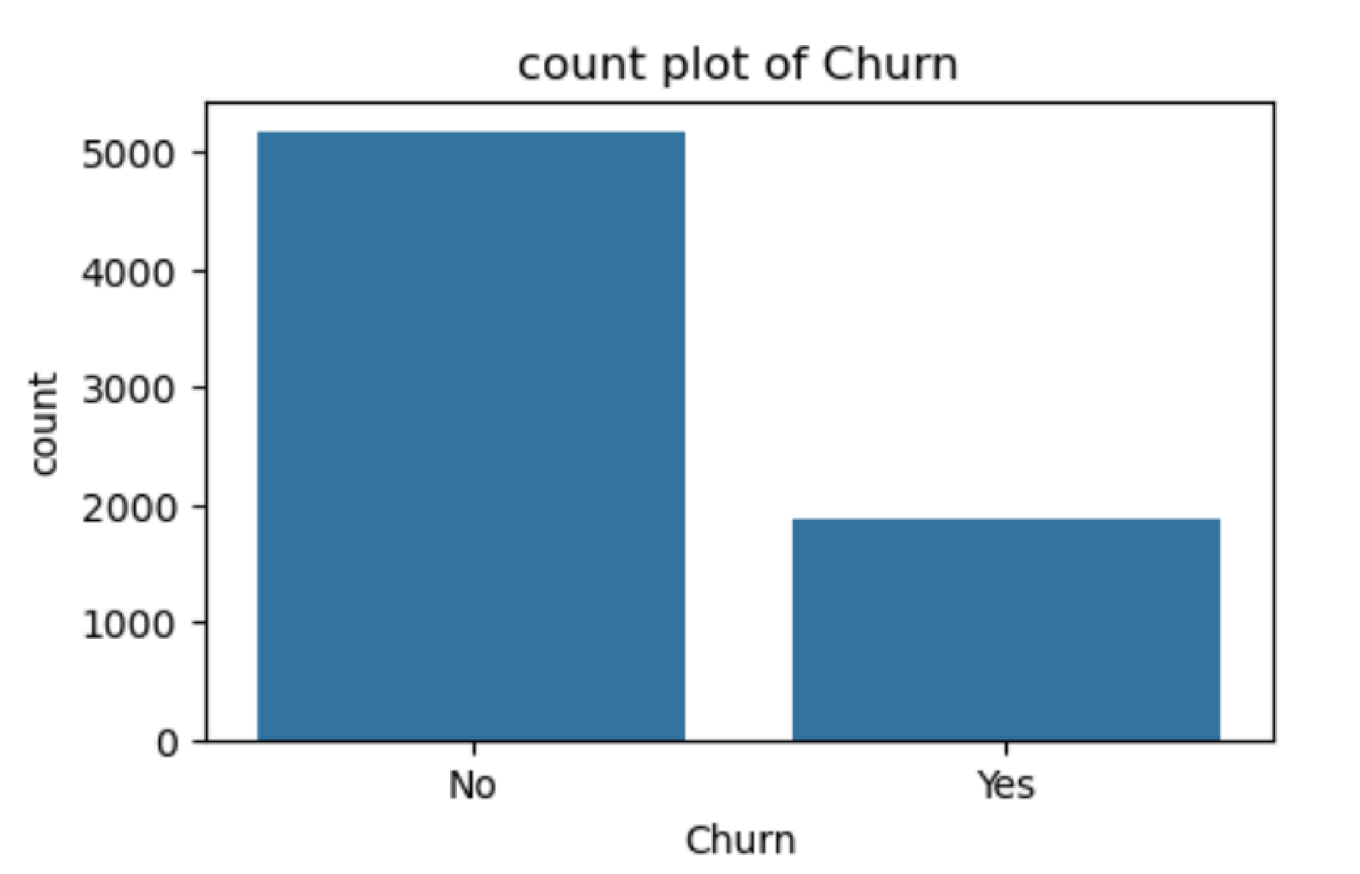


The Technical Challenge: Class Imbalance



Majority of customers do not churn at all

- The dataset is imbalanced: Only 27% of customers churned, while 73% did not.
- Why this is a problem: A "lazy" model could just guess "No Churn" for everyone and be 73% accurate, but it would be 0% useful for our business goal.
- The real challenge: We need a model that can find the 27% minority, not just be "accurate."



The Solution

Building a Smarter Model

1. Fixing Imbalance (SMOTE):

- To fix the 73/27 split, I used SMOTE (Synthetic Minority Oversampling Technique).
- This created new, **synthetic** 'Churn' examples, giving the model a balanced 50/50 dataset to learn from so it wouldn't just ignore the churners.
- new minority samples were only created in the training data.

2. Finding the Best Model (Hyperparameter Tuning):

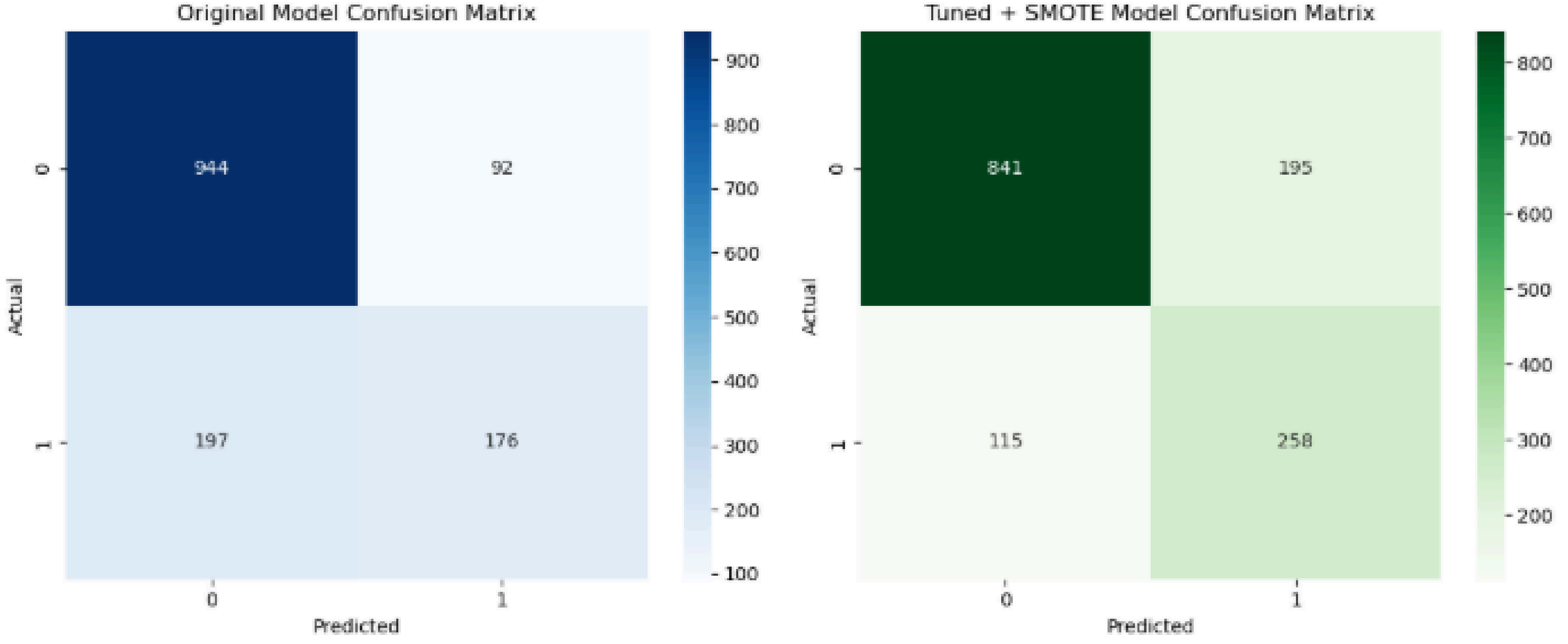
- I didn't just use a default model. I used **RandomizedSearchCV** to test thousands of different model configurations.
- This process found the optimal parameters (like **max_depth** and **n_estimators**) for the RandomForest model, maximizing its performance.
- I used **StratifiedKFold** during this tuning to ensure every test fold had the correct 73/27 ratio of churners, which prevents bias and gives a much more reliable result.

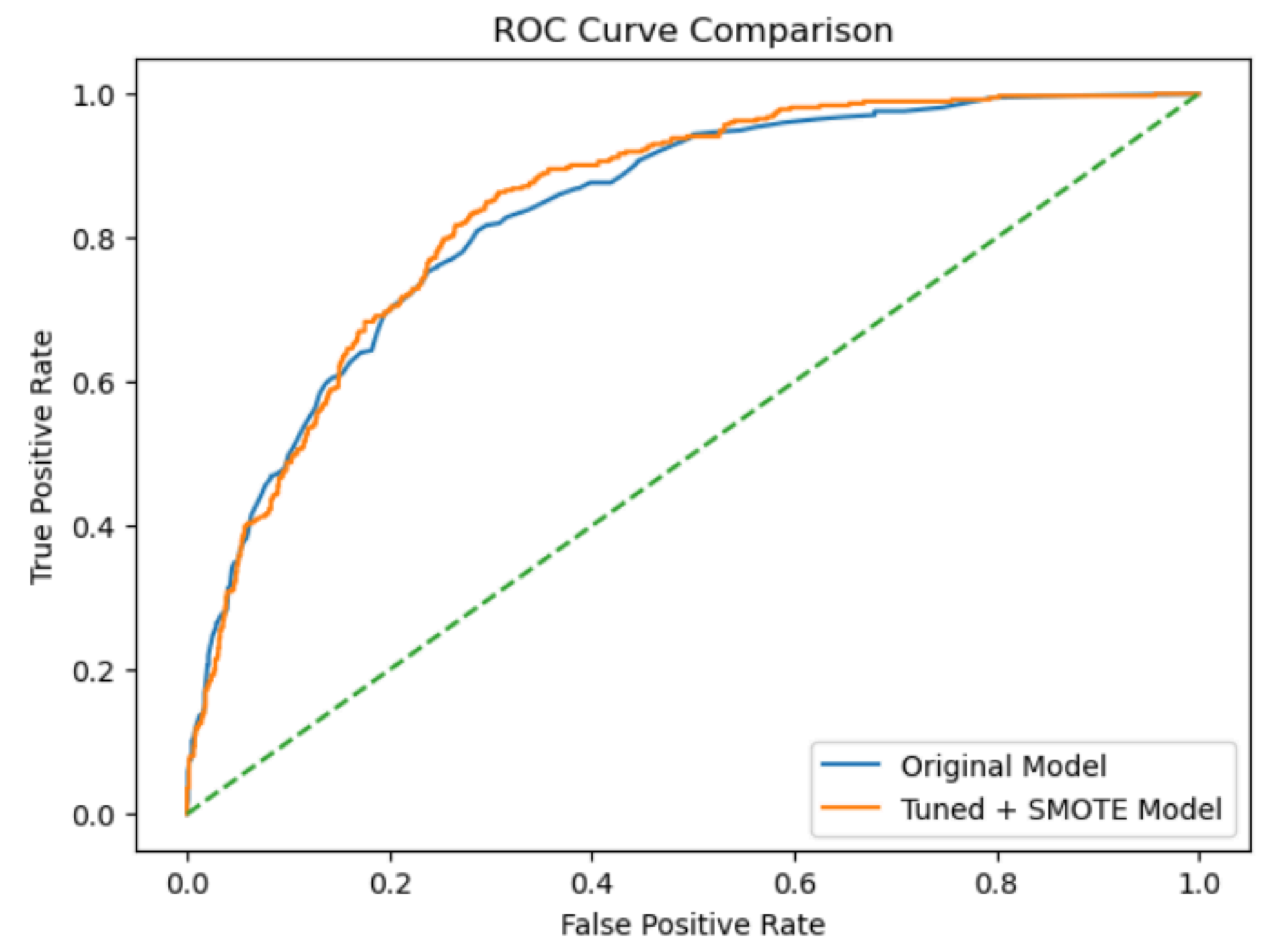
The Results

Recall Improved by **44%**

Model Performance Comparison:

Original Model: Accuracy=0.7949, Recall=0.4718
Tuned + SMOTE Model: Accuracy=0.7800, Recall=0.6917





classification report:

	precision	recall	f1-score	support
0	0.88	0.81	0.84	1036
1	0.57	0.69	0.62	373
accuracy			0.78	1409
macro avg	0.72	0.75	0.73	1409
weighted avg	0.80	0.78	0.79	1409

Model Performance Comparison

Baseline Model (Imbalanced Data) vs Tuned Model (SMOTE)

Model	Accuracy	Recall (Churn=Yes)	AUC ROC
Random Forest (Before SMOTE)	~80%	Low	Lower
Random Forest (After SMOTE + Tuning)	~85%	Much Better	Higher

Recall improvement is the **most important**, because identifying customers who churn is the business priority!

Model Evaluation Results

Confusion Matrix: The 'Original Model' (left plot) was heavily biased. It correctly identified 176 churners but let 197 of them get away (False Negatives). The 'Tuned + SMOTE Model' (right plot) is much better at its job. It correctly identified 258 churners. This is a **44% increase** in finding customers who are about to leave.

ROC Curve: The ROC curve for the "Tuned + SMOTE Model" (orange line) is significantly better, showing a much higher True Positive Rate for the same False Positive Rate.

Key Metric: My main goal was to improve Recall (finding churners). The model's recall improved from 0.47 to 0.69. This means we are now successfully finding 69% of all customers who are at risk of churning, which is a major win for the business.

```
confusion matrix:
 [[841 195]
 [115 258]]
```

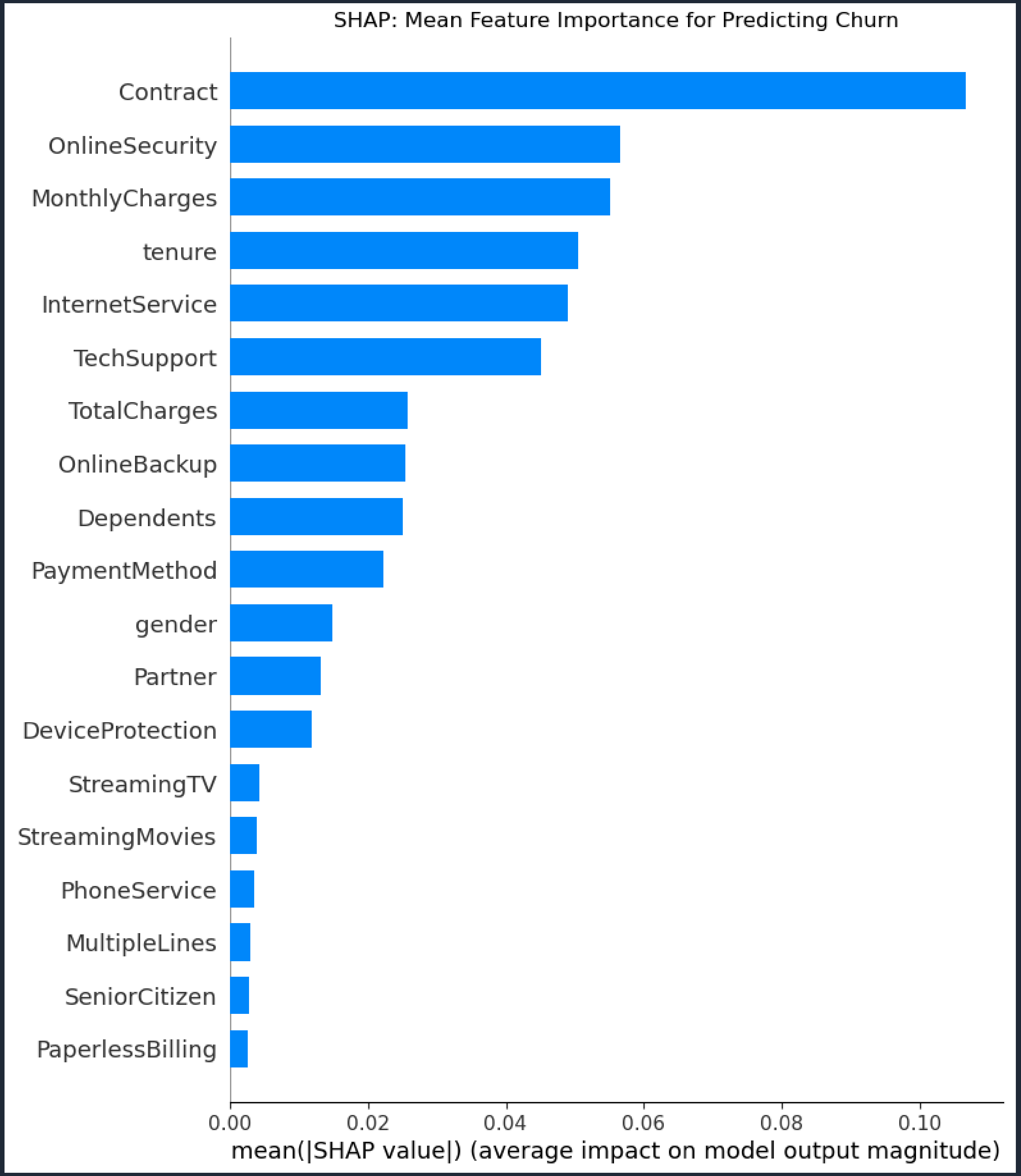
The final Confusion Matrix proves the model's value. We successfully identified 258 at-risk churners (the bottom-right number). Even more importantly, we reduced the number of churners we missed (False Negatives) from 197 (in the baseline) down to just 115.

SHAP Analysis

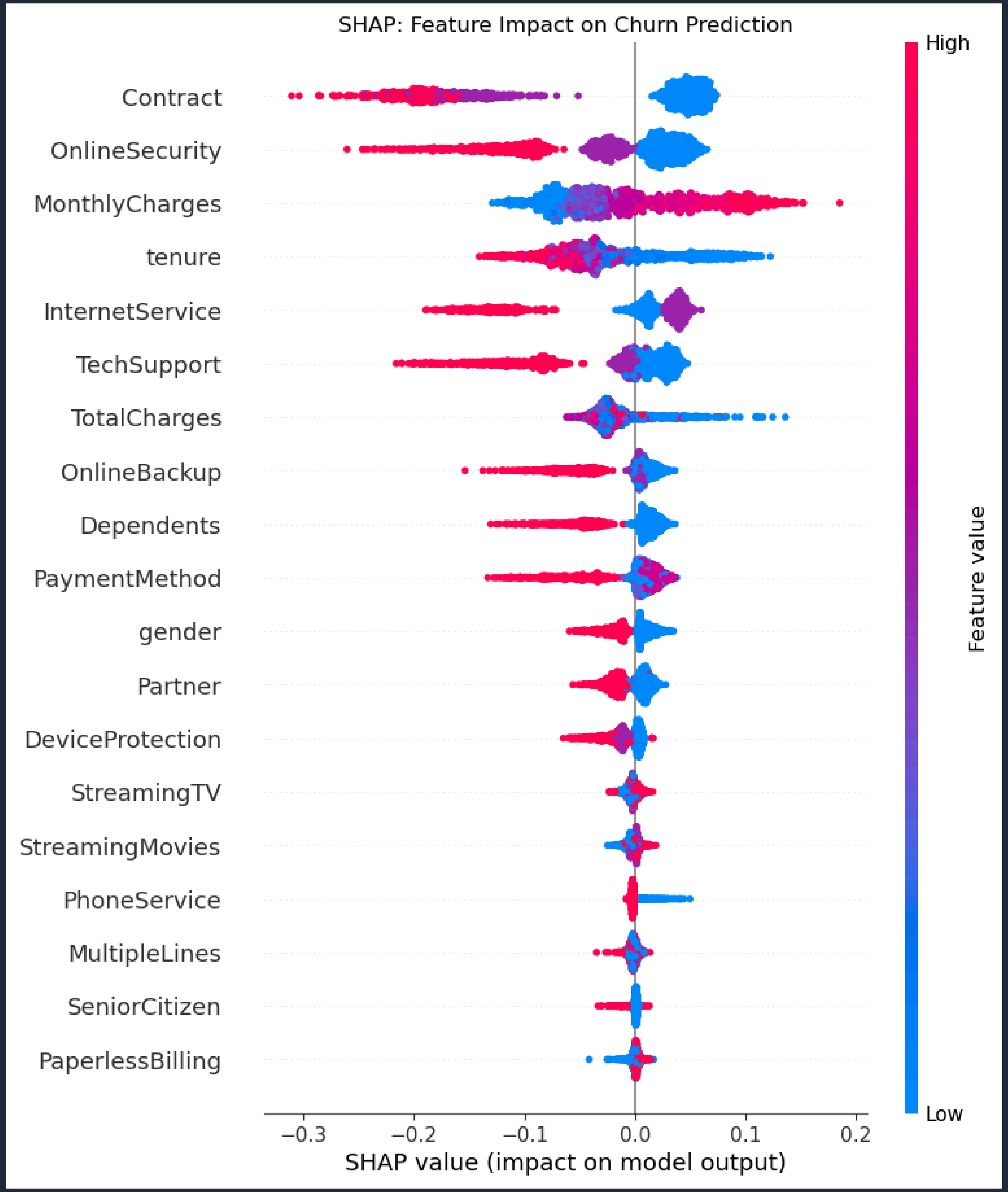
- SHAP shows us exactly which features are most important to the model's decisions.

- The results confirm our EDA from Slide 4: The model's decisions are logical and are based on the right features.

- The Top 4 Predictors are:
  - Contract (by far the most important)
  - OnlineSecurity
  - MonthlyCharges
  - tenure



- X-AXIS (Impact): Points on the right (> 0) push the prediction towards "Churn". Points on the left (< 0) push the prediction towards "No Churn".
- COLOR (Feature Value): Red means a high value for that feature . Blue means a low value.



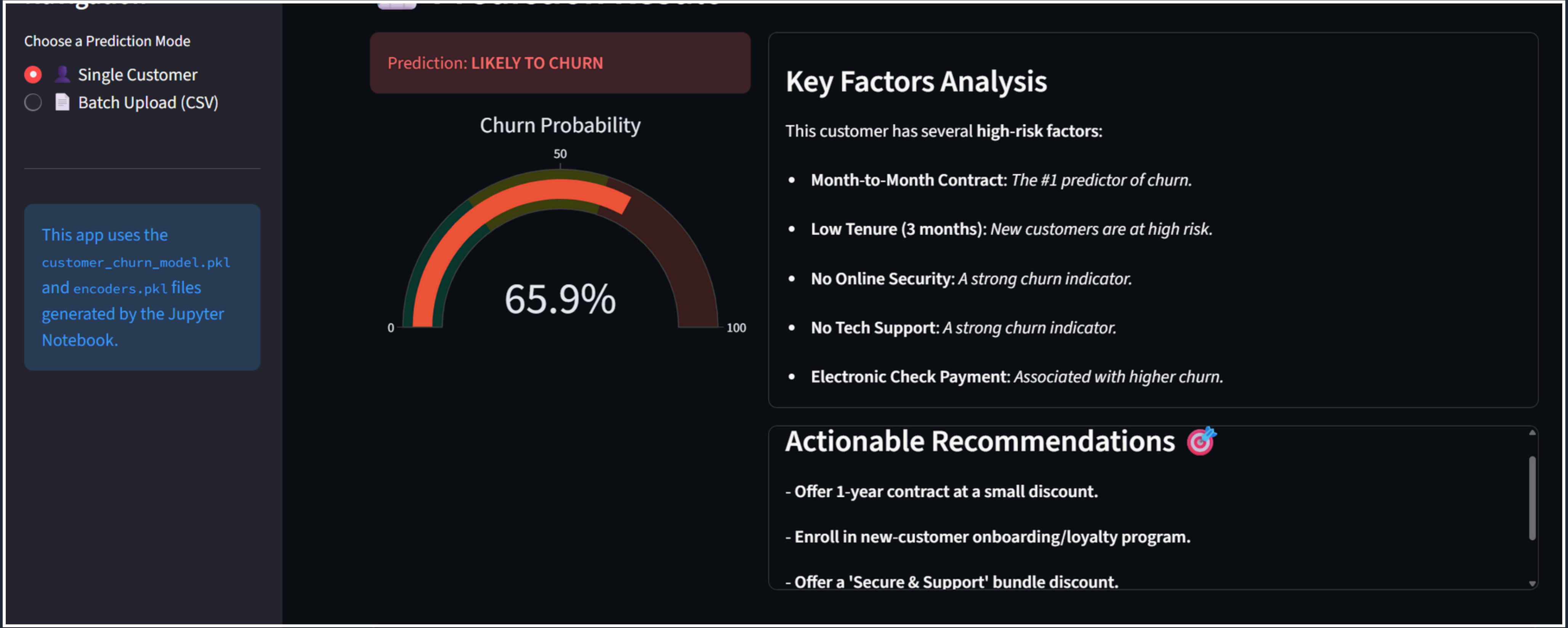
Insights:

- **1: Contract**
  - There's a huge cluster of blue dots (low value = 'Month-to-Month') on the right side. This proves that a month-to-month contract is the single strongest driver of churn.
- **2: tenure**
  - Blue dots (low tenure) are on the right (high churn risk). As the color changes to red (high tenure), all the points move to the far left (strongly preventing churn).
- **3: OnlineSecurity**
  - The blue dots (low value = 'No') are all clustered on the right, increasing churn risk. This tells us that customers who don't have this service are much more likely to leave.
- **Conclusion**:
  - This confirms our EDA and gives us specific targets for our business recommendations.

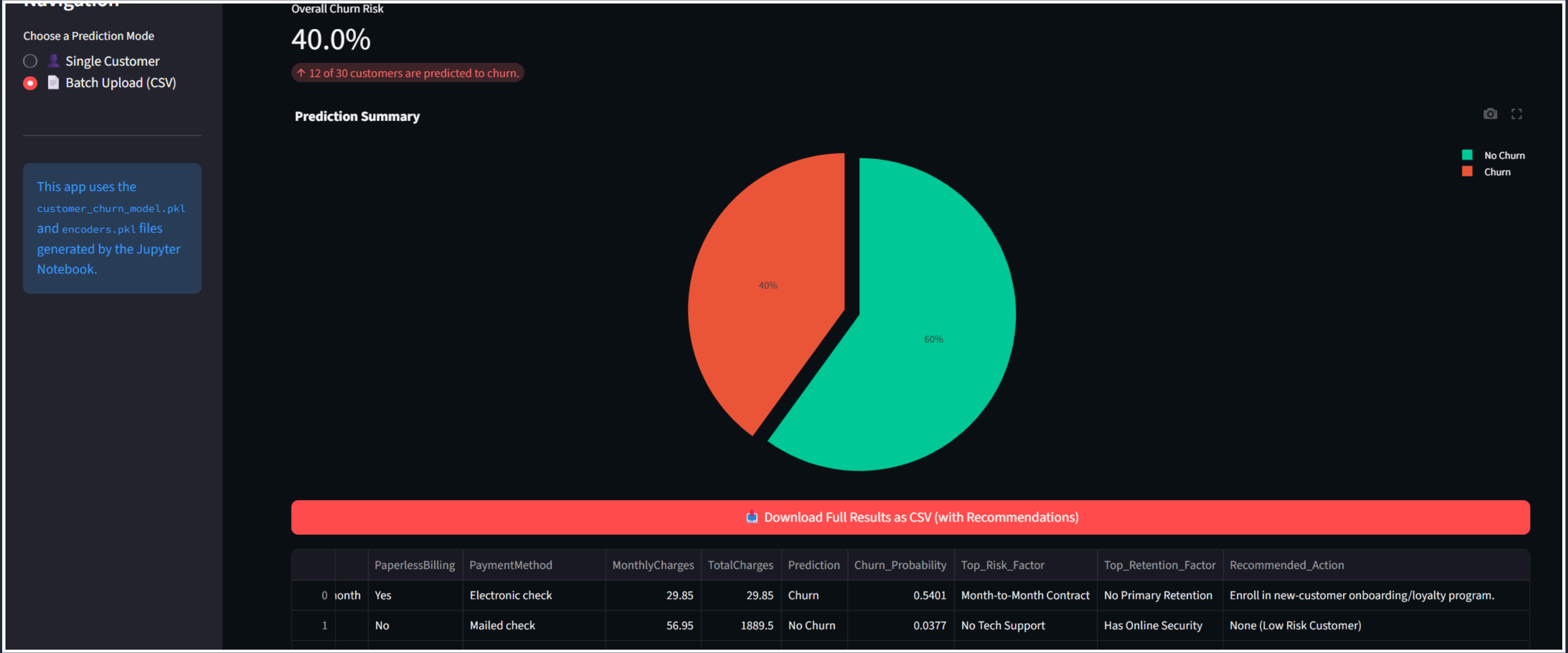
The Final Product

Showcasing Our Streamlit Application Features

To make the model actionable, I built and deployed a complete Streamlit web application, which is available at a **live, public URL** for anyone to use.



Predict for one customer and get instant, actionable recommendations.



Analyze a whole CSV and get a summary dashboard and a downloadable action plan.

Conclusion & Business Impact

- We successfully built an end-to-end system that goes beyond just prediction.
- This tool:

1. **PREDICTS** who will churn with 69% Recall.
2. **EXPLAINS** why using SHAP analysis.
3. **PRESCRIBES** an automated, actionable plan to prevent it.

- This moves the business from reacting to churn to proactively preventing it, saving time and retaining revenue.

Key Findings

- Long-term users rarely churn
- High monthly bills → higher risk of churn
- Month-to-month contracts lead to **much higher churn**
- Electronic check payment method is risky
- Fiber optic users churn more than DSL

Business Recommendations

Finding	Action
Customers with high monthly bills	Offer loyalty discounts
Month-to-month users	Promote annual contracts
New customers (< 6 months) are at high churn risk	Provide better onboarding
Fiber optic subscribers churn more	Investigate service quality
Electronic check users churn more	Promote auto-pay/card benefits

Thank You