Lab M1.03 - sklearn Model Training + Evaluation

Dina Bosma-Buczynska | 07.02.2026

Customer Churn Prediction with KNN

---

1. What is the Model's Accuracy?
Initial Model (K=5):

- **Test Accuracy:** 74.63%
- **Precision** 52.81%
- **Recall** 42.78%

Best Model After K Experimentation:

- **Best K Value:** 11
- **Test Accuracy:** 0.7690

  This is a **2.27 percentage point improvement** over K=5

Confusion Matrix:

```
=== CONFUSION MATRIX ===
                Predicted
              Retained  Churned
Actual Retained    890      143
       Churned     214      160
```

This means:
Total Test Samples: 890 + 143 + 214 + 160 = **1,407 customers**
Correct Predictions: 890 + 160 = **1,050 customers** (74.66% accuracy)

- 890 correctly identified as Retained
- 160 correctly identified as Churned

Incorrect Predictions: 143 + 214 = **357 customers** (25.34% error rate)

- 143 False Negatives (predicted Churned but actually Retained)
- 214 False Positives (predicted Retained but actually Churned)

Updated Performance Metrics (K=11):
Precision: 890 / (890 + 214) = **80.62%**

- When the model predicts "Retained," it's correct 80.62% of the time
- This is a **27.81 percentage point improvement** over K=5 (52.81%)

Recall: 890 / (890 + 143) = **86.16%**

- The model catches 86.16% of customers who actually stayed
- This is a **43.38 percentage point improvement** over K=5 (42.78%)

Specificity (Churn Detection Rate): 160 / (160 + 214) = **42.78%**

- The model only catches 42.78% of customers who actually churned
- This is the model's **weak point**

Comparison to Part 1:
The breast cancer model from Part 1 achieved approximately 90-95% accuracy, while this churn model achieved [YOUR]% accuracy. The churn prediction is more difficult because customer behavior involves personal choices, while medical measurements follow clearer patterns.

---

2. What features seem most important?
I did not perform feature importance analysis in this project. To answer this question properly, I would need to add code to:

- Compare churn rates across different categories (contract types, internet services, etc.)
- Calculate correlations between numeric features and churn
- Create visualizations showing feature distributions for churned vs. retained customers

  I decided to do this with some help ☺

What I do know from the preprocessing step:

- The dataset included **demographic features** (gender, SeniorCitizen, Partner, Dependents)
- **Service features** (PhoneService, InternetService, OnlineSecurity, TechSupport, etc.)

- **Account features** (Contract type, PaymentMethod, MonthlyCharges, TotalCharges, tenure) All of these were included in the model, but I did not measure their individual importance.
- **Churn by Contract Type** - Do month-to-month customers leave more?
- **Churn by Internet Service** - Does fiber optic have higher churn?
- **Monthly Charges** - Do churners pay more or less?
- **Tenure** - Are new customers or old customers more likely to leave?
- **Payment Method** - Does payment type matter?
- **Tech Support** - Does having support reduce churn?

Based on the feature exploration analysis, contract type is the strongest predictor of churn, with month-to-month customers churning at 42.7% compared to only 2.8% for two-year contracts (15 times higher risk).
The second most important feature is internet service type, where fiber optic customers churn at 41.9% versus 19.0% for DSL customers.
Other significant factors include payment method (electronic check users churn at 45.3% versus 15.3% for automatic payments), tech support status (41.6% churn without support versus 15.2% with support), monthly charges (churned customers pay $74.44 versus $61.31 for retained customers), and tenure (churned customers average 18 months versus 38 months for retained customers).
The data reveals that high-risk customers typically have month-to-month contracts, use fiber optic internet, pay manually, lack tech support, pay higher monthly fees, and are newer customers (under 2 years).

---

3. What Would You Recommend to the Company Based on Your Model?
Based on the Model's Performance:
With 76.9% accuracy, the model can help identify at-risk customers but is not accurate enough to use alone. The company should use it as one tool among many.

Understanding the Errors:
From the confusion matrix:
- False Positives (143 customers): The model predicted these customers would churn but they didn't. If the company offers them retention deals, it costs $50 per customer = $7,150 spent unnecessarily.
- False Negatives (214 customers): The model missed these churners. The company loses these customers and must spend $500 each to replace them = $107,000 in replacement costs.

Since replacement costs are much higher than retention costs, missing a churner is more expensive than offering an unnecessary deal.
Recommended Approach:
1. Use the model to identify potential churners.
2. Don't automatically send retention offers based solely on model predictions.
3. Have customer service review the list and add their knowledge about recent complaints, payment issues, service problems, and account changes.
4. Prioritize customers where both the model AND human judgment agree.

Based on feature analysis, to focus on month-to-month customers (42.7% churn), fiber optic users (41.9% churn), and electronic check payers (45.3% churn).

Pilot Program:
Start small to test if the model actually helps.
Select 200-300 customers the model predicts will churn, offer retention deals to half, don't contact the other half, and after 3 months compare how many actually churned in each group.

---

4. What are the limitations of your model?

Technical Limitations:
1. Moderate Accuracy: At 76.9% accuracy, the model makes mistakes on 23.1% of predictions. More critically, it only catches 42.8% of actual churners.
2. No Feature Scaling: KNN is sensitive to feature scales. Since I didn't normalize features, those with larger numbers (MonthlyCharges, TotalCharges) might dominate predictions unfairly.
3. Limited K Value Testing: I only tested K = 1, 3, 5, 7, 9, 11, 15. The best value might be between these numbers.

4. Single Train-Test Split: I used one 80-20 split. Cross-validation could provide more reliable accuracy estimates.

Data Limitations:

1. Historical Data Only: The model learned from past behavior. If customer preferences change, predictions become less accurate over time.
2. Missing Information: The model cannot see customer service quality, specific complaints, competitor offers, or life events that affect churn.
3. Dropped Missing Values: I removed customers with missing data, which may have created a biased sample.

Algorithm Limitations:

1. KNN Assumes Similarity Equals Same Behavior: Similar customers don't always make the same choices.
2. No Explanation: KNN doesn't explain why it made a prediction, so the company can't understand what causes churn.
3. Correlation Not Causation: Month-to-month contracts correlate with churn, but customers might choose month-to-month because they already plan to leave.

CONCLUSION

The KNN model achieved 76.9% accuracy (with K=11) in predicting customer churn. While this is lower than the 90-95% accuracy from the cancer prediction model, it demonstrates that machine learning can find patterns in customer behavior.

The model has value as a screening tool to identify potentially at-risk customers, but it should not make decisions automatically. The company should combine model predictions with human judgment, customer service data, and business knowledge.

Most importantly, the model's effectiveness should be tested with a pilot program before full deployment. Only real-world results will show whether the predictions actually help retain customers.

Key Learning: Predicting human behavior (churn) is harder than predicting physical measurements (cancer) because people's choices are influenced by many factors that cannot be captured in a dataset.
