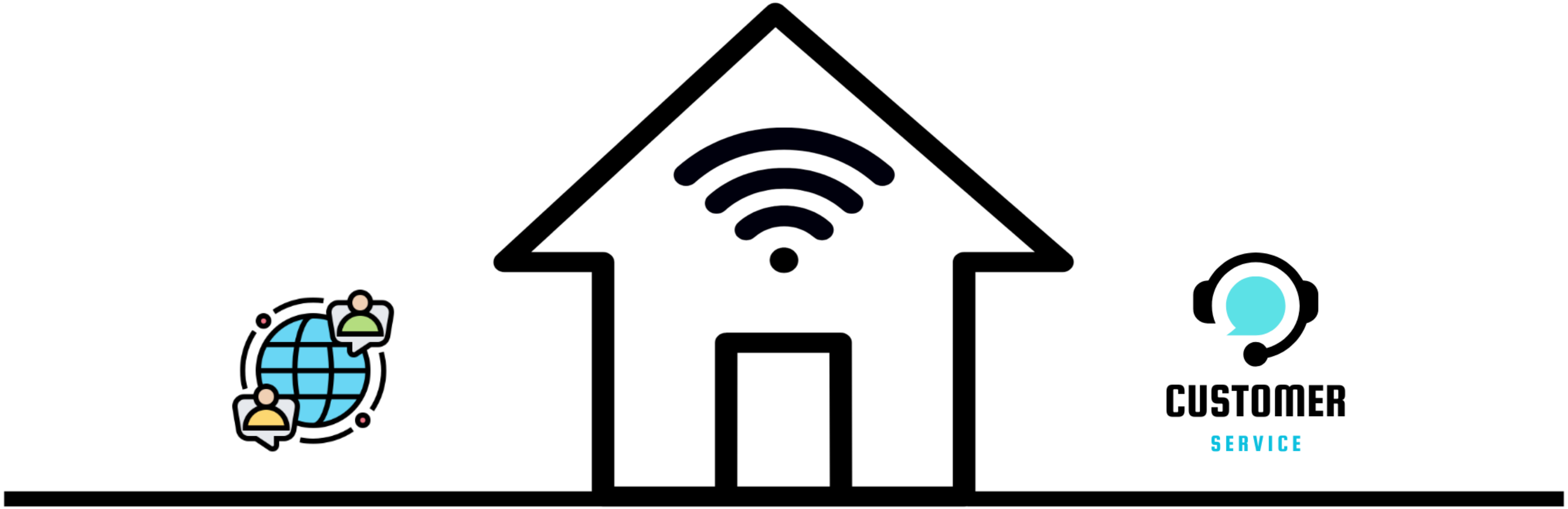
UNIVERSITÄT MANNHEIM

Telco Customer Churn Project

Data Mining I Project



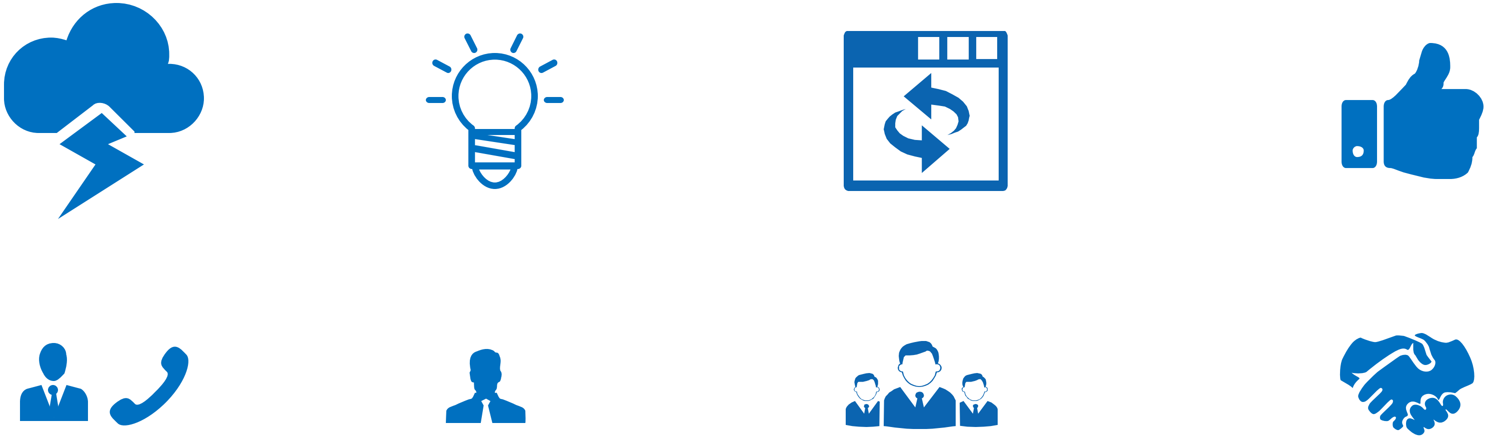
Team 9: Zhiqi Yang, Elise Wolf, Xi Liu, Shiqi Zhou, Yenan Chen

02.12.2024

Overview

There goes a scenario:

One day, a customer called to complain, let’s see how our model helps retain this customer…



Goals and Approach

UNIVERSITÄT MANNHEIM

Telecom customer management grows tougher

External: Turbulent and shifting industry

- Intensified competition
- Shifting consumer behavior

Internal: CRM capability challenges

- Action effectiveness
- Resource drain issues
- Customers acquisition challenge

Core Issue

Acquiring new customers tends to be **far more expensive** than retaining existing one

Goals (What)

Empower telecoms with predictive insights for **early action** and **stable customer retention**

Actions (How)

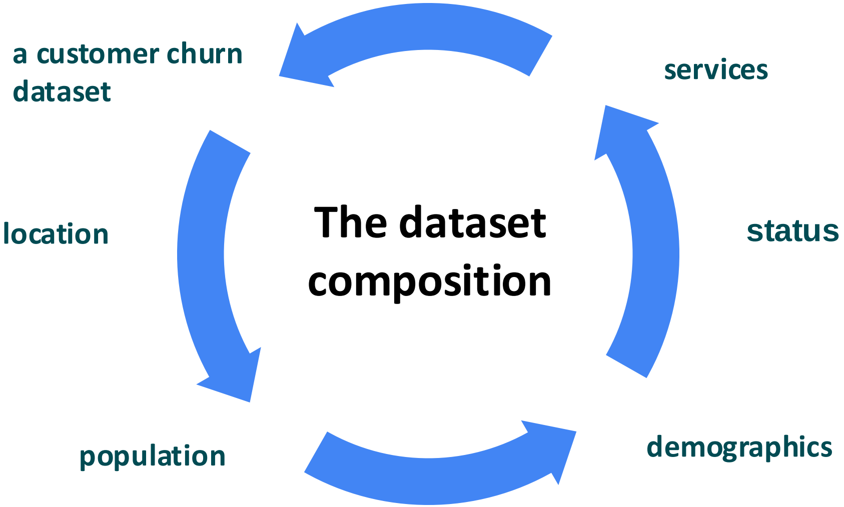
Develop Model → Apply ML → Implement Strategies

Dataset Structure

91 features
*after splitting dataset

67 Categorical

24 continuous

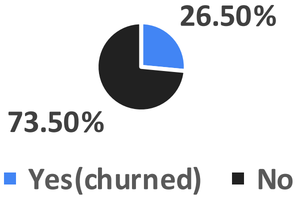


Dataset overview

- **7,000+**customer records
- The features spanning both categorical and numerical variables

Dataset balance

- The target variable, Churn, **is imbalanced**
- 26.5% labeled as Yes (churned) and 73.5% as No.



*This imbalance will be careful handling during model training, using techniques to ensure predictive fairness

1. Introduction

Hello, how can I help you?

Can you tell me your CustomerID?

UNIVERSITÄT MANNHEIM

Telco Customer Churn Dashboard

1875-QIVME | Search

What do we know about you?

Telco Customer Churn Dashboard

1875-QIVME | Search

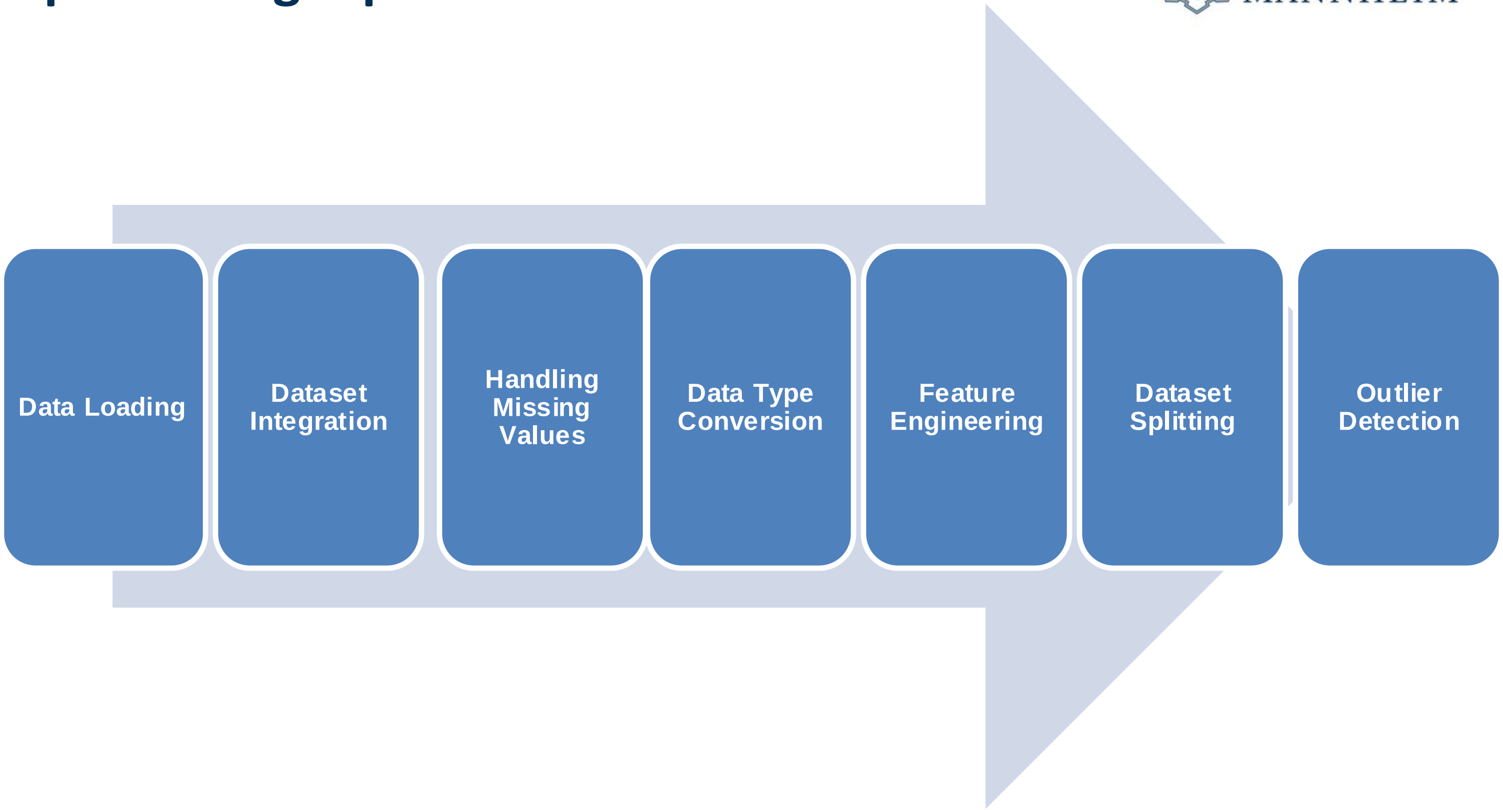
Customer ID: 1875-QIVME

Churn Prediction: Likely to Churn

UNIVERSITÄT MANNHEIM

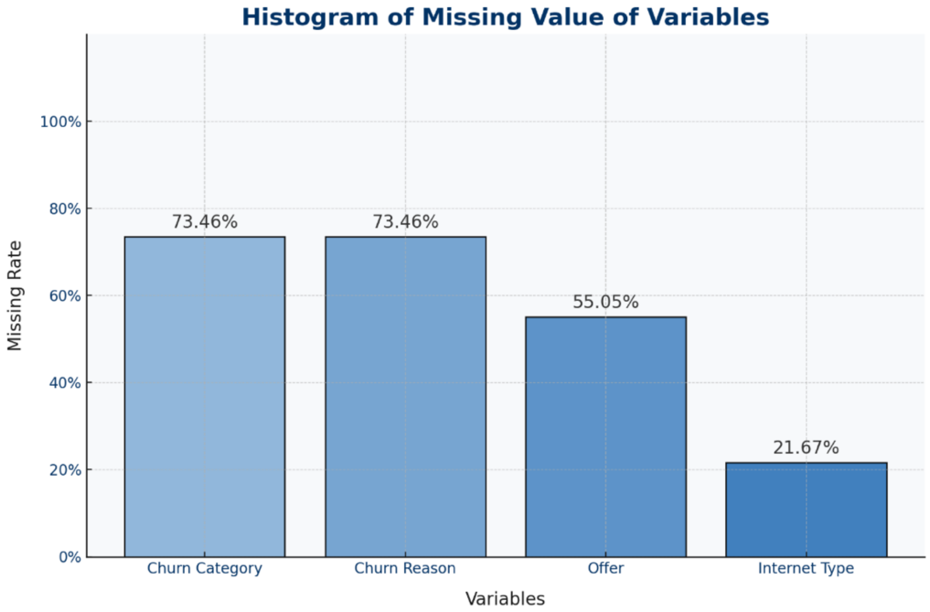
2. Preprocessing

Preprocessing Pipeline



Preprocessing

- **Handling Missing Values**
  - Significant Missing Values: substantial portion of the dataset affected, suggesting a strong relationship between the two columns.
  - Explanation: align with rows where the ‘Churn’ column is ‘No’, implying that these details were not recorded because they are not applicable to non-churning customers.



Preprocessing: Data Type Conversion

- **Removal of Unnecessary Columns**:

  - Removed columns with entirely unique entries (e.g., IDs) to retain only meaningful features.

- **Categorical Variable Encoding**:

  - Label Encoding: Applied to binary variables like "Gender" and "Senior Citizen."

  - One-Hot Encoding: Used for features with multiple categories

- **Post-Encoding Compatibility Check**:

  - Ensured all columns were numeric for compatibility with machine learning algorithms.

Preprocessing: Feature Engineering

- **Key Approaches**
  - **Interaction Features**:
    - Combine variables (e.g., tenure vs. age) to reveal trends
  - **Aggregation**:
    - Summarize binary features (e.g., total services subscribed)
  - **Group-Based Features**:
    - Segment customers (e.g., revenue tiers by charges)
  - **Domain Transformations**:
    - Align features with business metrics (e.g., annualized charges, refund-to-charges ratio)

Preprocessing

- **Data Splitting**
  - Split the data into training and test sets:
    - 80% (5,634 rows) for training
    - 20% (1,409 rows) for testing
- **Outlier Detection**
  - Outlier detection and removal are performed only on the training set
  - The data is scaled when applying each method
  - 3,885 rows remain in the training set after outlier removal

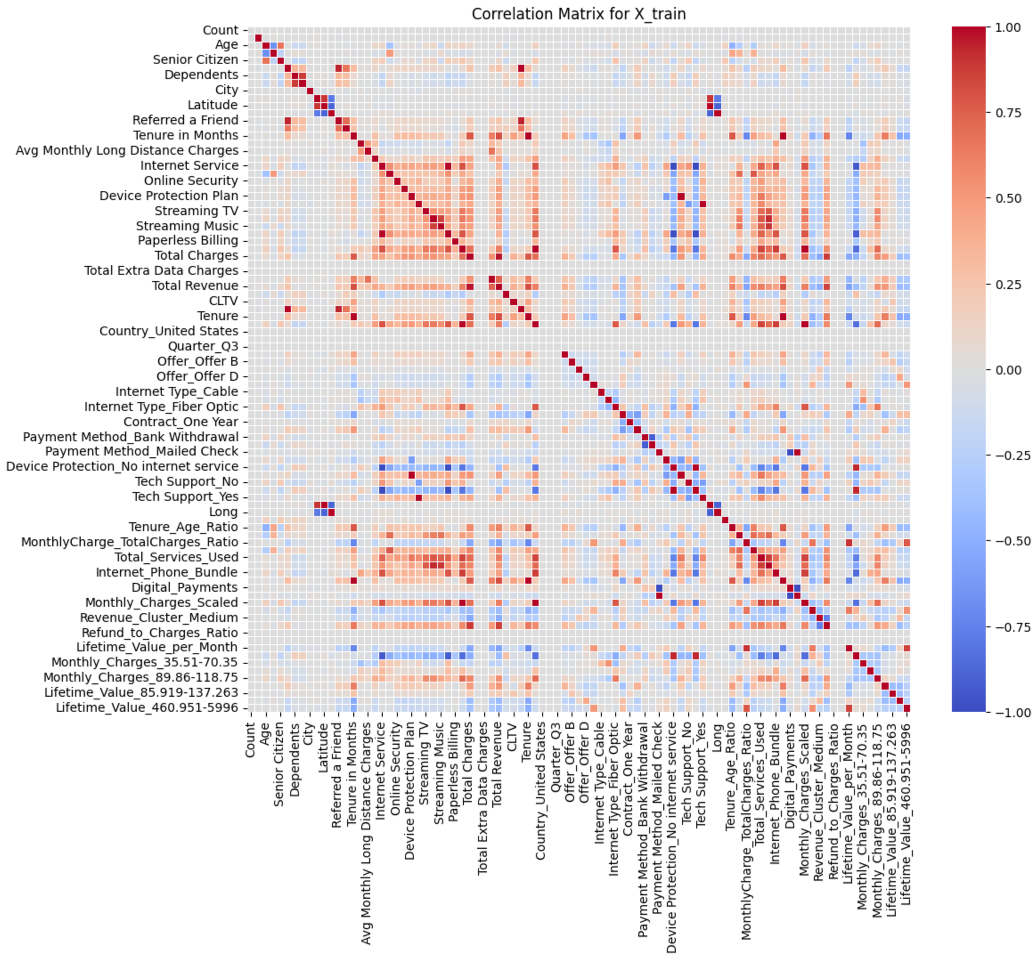
3. Clustering

Correlation Analysis

Long
Longitude:
1.0

Monthly Charges
Monthly_Charges_Scaled:
1.0

Tenure
Tenure_in_Years:
0.99982

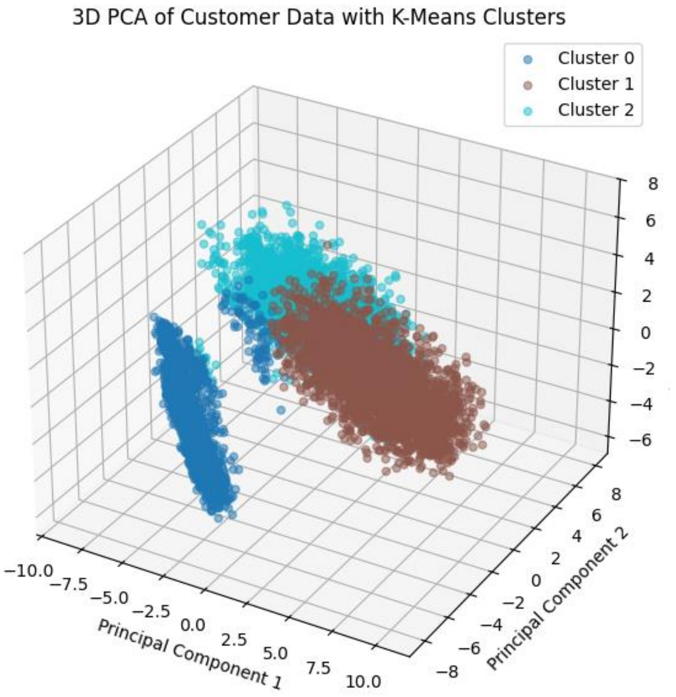


Team 9: Zhiqi Yang, Elise Wolf, Xi Liu, Shiqi Zhou, Yenan Chen
02.12.2024

Finding common Customer Profiles

Differentiators for the Churn Value – PCAs:
Loyality, Total Revenue, Total Charges

Cluster 0 (Churn 0.03)	Moderate revenue, low churn, **stable and satisfied**
Cluster 1 (Churn 0.08)	High revenue, low churn, **most valuable**
Cluster 2 (Churn 0.66)	Low revenue, high churn, **dissatisfied and at risk**



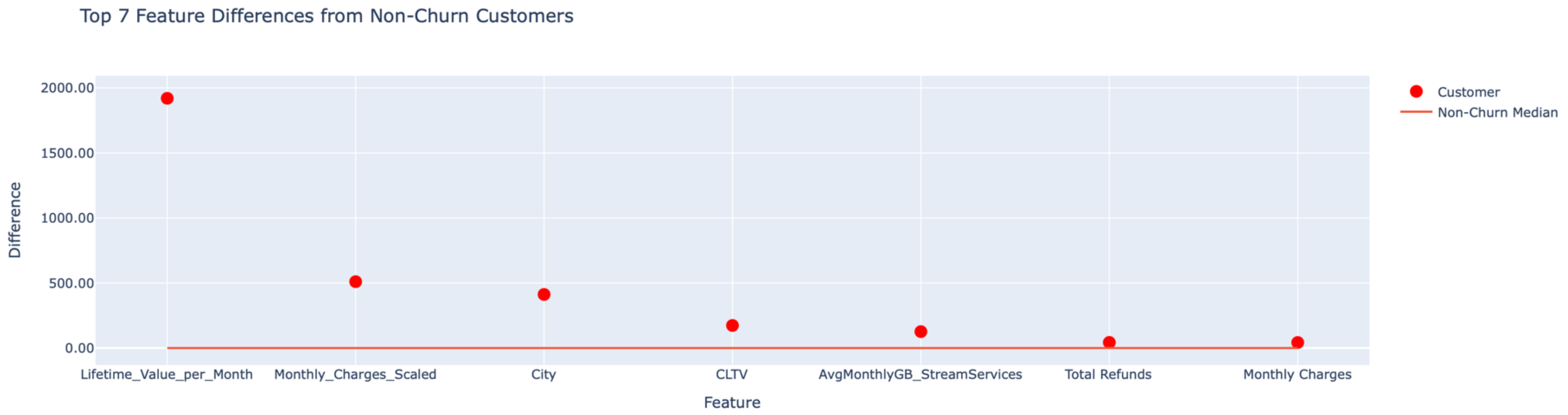
How we make data speak

Telco Customer Churn Dashboard

1875-QIVME Search

Customer ID: 1875-QIVME

Churn Prediction: Likely to Churn



4. Models Analysis

Different Approaches

Naive Bayes

Logistic Regression

KNN

Nearest Centroid

Decision Tree

Random Forest

XG Boost

SVM

MLP

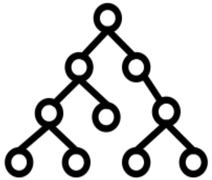
4 Main Types

Probabilistic Models

- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Logistic Regression

Distance-Based Models

- K-Nearest Neighbors (KNN)
- Nearest Centroid



Tree-Based Models

- Decision Tree
- Random Forest
- XG Boost

Optimization and Kernel-Based Models

- Support Vector Machine (SVM)
- Multilayer Perceptron (MLP)

Similar Procedures in 10 Models

I. Train Models

II. Hyperparameter Tuning

III. Cross-Validation

IV. Performance Metrics

Probabilistic Models - Naive Bayes

Gaussian Naive Bayes

I. **Feature Preparation**
   - Discretization of **continuous** features
II. Train Model
III. Hyperparameter Tuning
IV. Handle Data Imbalance

Multinomial Naive Bayes

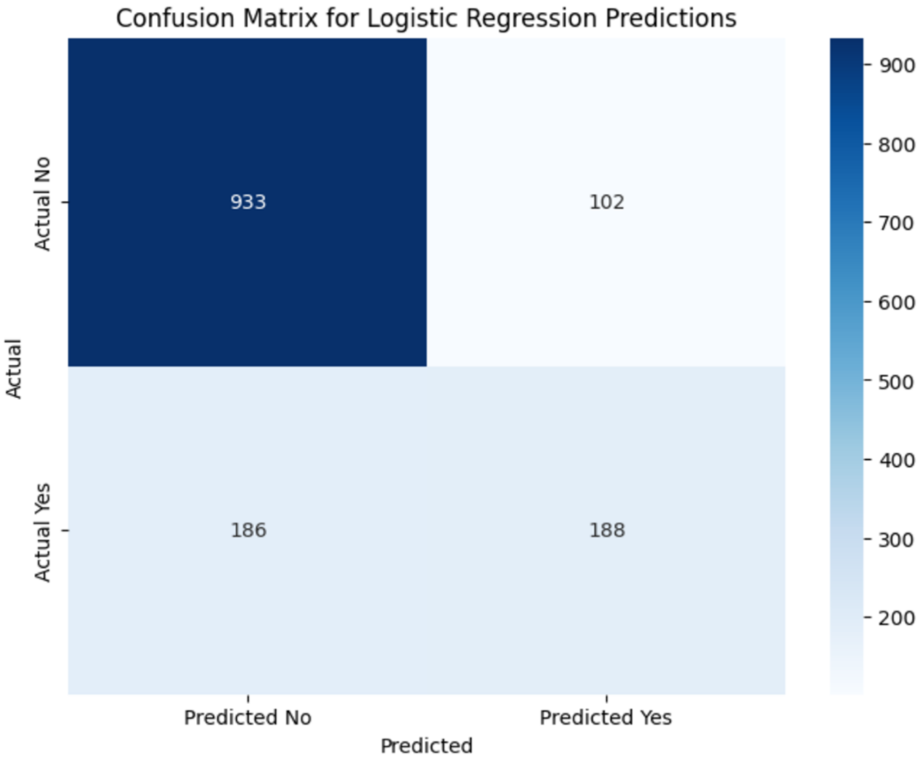
I. **Feature Preparation**
   - Discretization of **boolean-encoded** or **count-based** features
II. Train Model
III. Hyperparameter Tuning
IV. Handle Data Imbalance

Probabilistic Models - Logistic Regression

I. Initialize and Train Model

II. Cross-Validation

III. Predictions and Final Metrics



Distance-Based Models - KNN

UNIVERSITÄT MANNHEIM

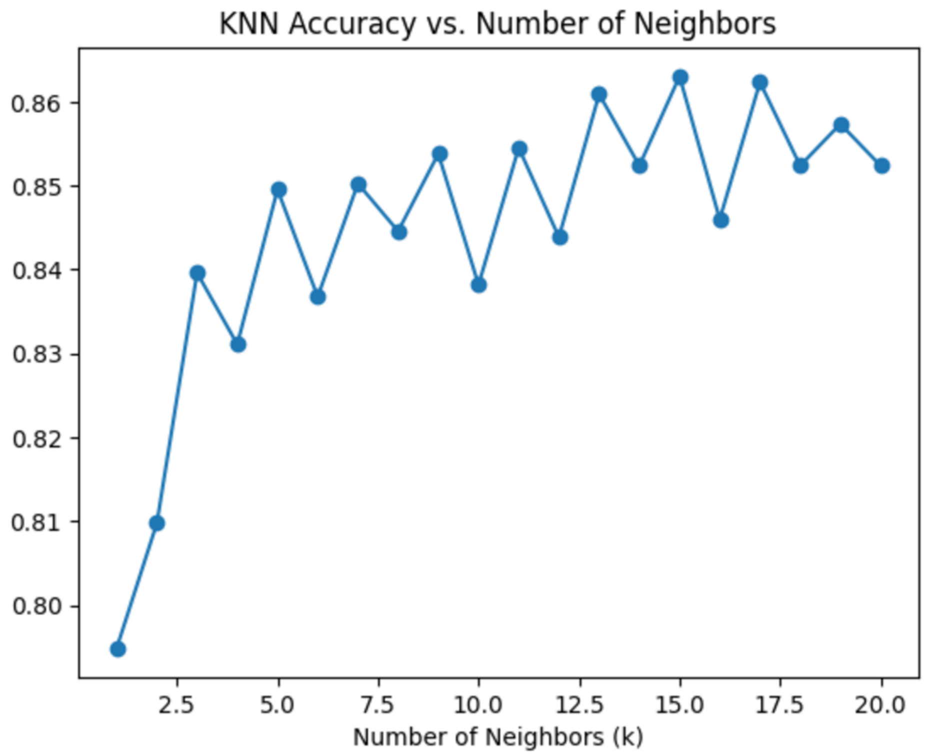
I. Standardization

II. Baseline KNN Model

III. Find the Optimal Number of Neighbors (k)

IV. Optimize KNN Model Training

V. Final Performance Metrics



Team 9: Zhiqi Yang, Elise Wolf, Xi Liu, Shiqi Zhou, Yenan Chen

02.12.2024

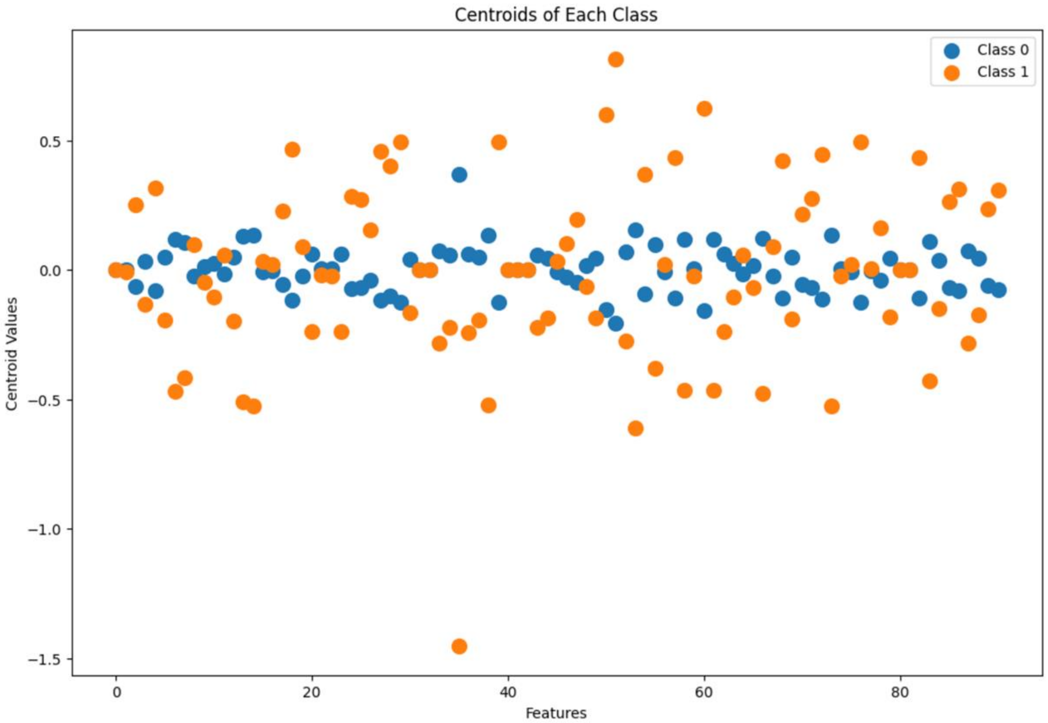
Distance-Based Models - Nearest Centroid

I. **Standardization**

II. Train Model

III. Cross-Validation

IV. Performance Metrics



Tree-Based Models
– Decision Tree & Random Forest & XG Boost

Decision Tree

I. Data Preparation

II. Train Model

- Create **entire** Tree

III. Cross-Validation

IV. Performance Metrics

Random Forest

I. Data Preparation

II. Train Model

- Train a **ensemble** of 100 smaller trees

III. Cross-Validation

IV. Performance Metrics

XG Boost

I. Data Preparation

II. Train Model

- Build trees **sequentially**

III. Cross-Validation

IV. Performance Metrics

Optimization and Kernel-Based Models - SVM & MLP

SVM

I. **Train Model**
   - Train using the fit() method on the **standardized data**
II. Cross-Validation
III. Performance Metrics

MLP

I. **Train Model**
   - Combine preprocessing and model training in a **Pipeline**, executed via fit()
II. Cross-Validation
III. Performance Metrics

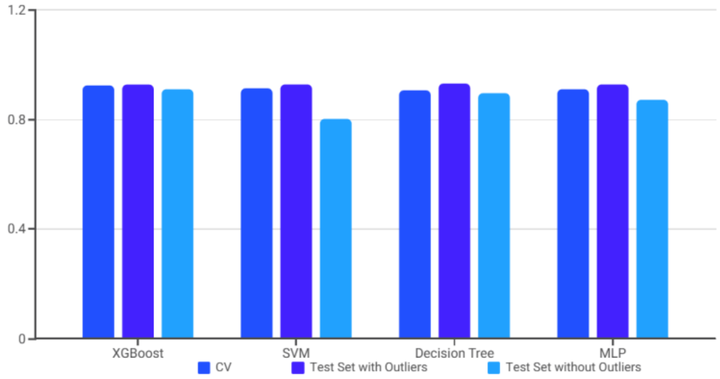
5. Evaluation

Evaluation Metrics

Measurement

- Baseline Model
  - Baseline included predicting the majority class and using rule-based heuristics.
  - Purpose: To understand how much value advanced models add.

- Evaluation
  - Cross Validation Performance
  - Test Set Performance
  - Outlier Remove



Evaluation Metrics

Final Results and Recommendation

- Evaluation based on F1-Score
- XG Boost performs best in Training
- Decision Tree performs best on Test set

Model	Accuracy	Precision	Recall	F1	ROC AUC
Decision Tree Prediction	0.963804	0.942466	0.919786	0.930988	0.949748
XGBoost Prediction	0.963804	0.965418	0.895722	0.929265	0.942064
SVM Prediction	0.963804	0.968116	0.893048	0.929068	0.941210

Model	Accuracy	Precision	Recall	F1	ROC AUC
XGBoost CV	0.959709	0.941276	0.905029	0.922662	0.990757
SVM CV	0.955985	0.951584	0.878962	0.913684	0.987776
MLP CV	0.952608	0.916340	0.904349	0.910133	0.989161
Decision Tree CV	0.951188	0.932193	0.881620	0.905068	0.977231
Random Forest CV	0.938231	0.951006	0.809351	0.874159	0.968387
Nearest Centroid CV	0.791802	0.830656	0.791802	0.801299	NaN
MNB CV	0.789140	0.825828	0.789140	0.798502	0.869418
GNB CV	0.854992	0.686477	0.836166	0.753798	0.908905
KNN CV	0.867953	0.782722	0.696398	0.736450	0.910824
Baseline Rule Based CV	0.612373	0.791717	0.612373	0.629931	0.704378
Baseline CV	0.734647	0.539706	0.734647	0.622266	NaN
Baseline Random CV	0.611108	0.613337	0.618740	0.600410	NaN
Logistic Regression CV	0.790561	0.661667	0.432828	0.522673	0.812847

Feature Importance

Customer Satisfaction is a Primary Driver

The Satisfaction Score stands out as the most significant predictor of customer churn in both SVM and XGBoost models. This aligns with intuition, as customers with low satisfaction are more likely to leave.

Contract Type Reflects Commitment

Contract types, such as month-to-month and two-year, are highly influential. Month-to-month contracts are associated with higher churn rates while longer-term contracts indicate greater customer commitment.

Engagement is Crucial

Customer engagement is a strong indicator of loyalty. Features like Number of Referrals and Online Security are highly influential. Customers who are actively engaged with the company are less likely to churn.

Feature Contributions Vary

The feature importance analysis highlights that different models can assign varying levels of importance to specific features. This emphasizes the need for a holistic understanding of feature contributions across models.

Preferred Model

Dashboard

- Predictive performance.
- Interpretability of feature importance.
- Alignment with the dashboard's functionality requirements.

Preferred Model-XGBoost

- Identify critical features
- Tailor retention strategies
- Enhance customer interactions with precise, data-driven insights.

Future Work

Time Series Analysis

Improve churn prediction by incorporating temporal trends in customer behavior. This includes features like changes in usage patterns or payment histories.

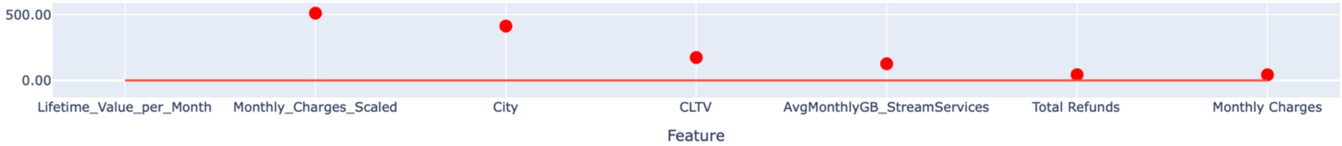
Real-Time Prediction

Integrate real-time data streams into the prediction framework. This would allow for more timely interventions based on changes in customer behavior.

Dashboard Expansion

Enhance the customer retention dashboard with interactive features. This allows customer service representatives to simulate different retention strategies.

The Power of Insight: Personalized Recommendations



Recommendations to Retain Customer:

Customer Satisfaction

- Low satisfaction. Offer personalized support.

Security and Support

- Low satisfaction. Offer personalized support.
- No online security. Highlight benefits.

Contracts and Payments

- Month-to-month contract. Promote long-term benefits.
- Not on two-year contract. Discuss perks.
- No credit card payment. Recommend for convenience.

Streaming and Services

- Young customer. Highlight appealing services.

Referrals and Offers

- Low satisfaction. Offer personalized support.

UNIVERSITÄT MANNHEIM

Great, to keep you as a customer!

Thank you for your attention!

github.com/eelisee/telco-customer-churn

Team 9: Zhiqi Yang, Elise Wolf, Xi Liu, Shiqi Zhou, Yenan Chen