**Customer Churn Prediction: High Dimensional Statistical Analysis**

**Executive Summary**:  
This analysis of the Telco Customer Churn dataset employs Principal Component Analysis (PCA) and classification techniques to uncover drivers of customer churn, informing retention strategies. PCA reveals that MonthlyCharges, TotalCharges, and tenure account for 39.21% of variance in PC1, with tenure and contract type driving PC2 (18.68%). Logistic Regression achieved the highest predictive accuracy (0.82), followed by Random Forest (0.79) and K-Nearest Neighbors (KNN, 0.80). Random Forest highlights TotalCharges, tenure, and MonthlyCharges as key predictors. High-risk groups include customers with month-to-month contracts (42.71% churn), fiber optic services (41.89%), and electronic check payments (45.29%). Recommended actions include loyalty programs, long-term contract incentives, fiber optic service enhancements, and payment method improvements to reduce churn and enhance financial stability.

**Introduction**

Customer churn, the phenomenon of customers discontinuing services, poses a significant challenge for telecommunications companies, impacting revenue and market competitiveness. The Telco Customer Churn dataset, comprising 7,043 customer records with 21 features—spanning demographics (e.g., gender, senior citizen status), service usage (e.g., internet service type), billing details (e.g., MonthlyCharges, TotalCharges), and a binary churn label (26.54% churn rate)—provides a rich resource for analyzing churn drivers. The objective of this project is to develop predictive models and identify actionable insights to inform retention strategies, addressing a critical business need in a high-dimensional data context.

Exploratory Data Analysis (EDA) uncovers pronounced patterns in churn behavior. Customers with month-to-month contracts exhibit a 42.71% churn rate, significantly higher than those with one-year (11.27%) or two-year contracts (2.83%), suggesting contract duration as a pivotal factor (Figure 1). Fiber optic service users show a 41.89% churn rate, compared to 18.96% for DSL and 7.40% for no internet service, indicating service-specific dissatisfaction. Payment methods also influence churn, with electronic check users at 45.29% versus 15.24% for credit card (automatic) users. Senior citizens churn at 41.68%, nearly double the 23.61% rate for non-seniors. Numerical features reveal further insights: tenure averages 32.37 months (standardized mean: 0.00, Table 1), with a skew toward shorter durations, while MonthlyCharges and TotalCharges vary widely, reflecting diverse billing profiles. These findings underscore the need for advanced statistical methods to handle high-dimensional data and extract meaningful patterns.

This report leverages Principal Component Analysis (PCA) to reduce dimensionality and uncover latent structures, and applies classification techniques—K-Nearest Neighbors (KNN), Random Forest, and Logistic Regression—to predict churn. By integrating these methods, the analysis aims to provide robust predictions and actionable business recommendations, contributing to strategic decision-making for customer retention.

**Table 1: Summary Statistics for Numerical Features**

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Mean | Std Dev | Range |
| tenure | 0.00 | 1.00 | [-1.32, 1.61] |
| MonthlyCharges | 0.00 | 1.00 | [-1.55, 1.79] |
| TotalCharges | 0.00 | 1.00 | [-0.99, 2.83] |

**Figure 1: Churn Rate by Contract Type**  
A graph of a chart

AI-generated content may be incorrect.

**Methods and Results**

**Principal Component Analysis**

High-dimensional datasets, like the Telco dataset with 30 preprocessed features (numerical features standardized, categorical features one-hot encoded), pose challenges for analysis due to multicollinearity and computational complexity. PCA addresses these by transforming features into orthogonal principal components that capture maximum variance. The dataset was preprocessed to standardize numerical features (tenure, MonthlyCharges, TotalCharges) and encode categorical variables (e.g., Contract, InternetService), resulting in a 30-dimensional feature space suitable for PCA.

The PCA results reveal significant structure. The first principal component (PC1) explains 39.21% of the total variance, driven predominantly by MonthlyCharges (loading: 0.502), TotalCharges (0.491), and tenure (0.333), indicating that billing-related attributes and customer longevity are central to data variability (Table 2). PC2 accounts for 18.68% of variance (cumulative: 57.89%), with strong contributions from tenure (0.615) and Contract\_Two year (0.221), reflecting contract duration as a secondary pattern. PC3 contributes 6.10% (cumulative: 63.99%), but subsequent components have diminishing returns, as shown in the scree plot (Figure 2), which suggests retaining the first two components for interpretation. The biplot (Figure 3) visualizes customer distribution in the PC1-PC2 plane, with churned customers (red) clustering toward higher PC1 values (high charges, short tenure), and non-churned customers (blue) toward lower PC1 and higher PC2 (longer contracts), confirming distinct behavioral profiles.

PCA’s dimensionality reduction facilitates visualization and interpretation while preserving key patterns. The loadings (available in pca\_loadings.csv, appendix) provide detailed feature contributions, enabling targeted business insights.

**Table 2: Proportion of Variance Explained by PCA**

|  |  |  |
| --- | --- | --- |
| Component | PVE (%) | Cumulative PVE (%) |
| PC1 | 39.21 | 39.21 |
| PC2 | 18.68 | 57.89 |
| PC3 | 6.10 | 63.99 |

**Figure 2: Scree Plot**  
A graph with a line

AI-generated content may be incorrect.

**Figure 3: PCA Biplot: PC1 vs PC2**  
A graph of red and blue dots

AI-generated content may be incorrect.

**Classification**

To predict churn, three classification methods were applied: K-Nearest Neighbors (KNN), Random Forest, and Logistic Regression. The preprocessed dataset was split into 80% training (5,634 records) and 20% test (1,409 records) sets, ensuring robust model evaluation. Each method was implemented using Python’s scikit-learn library, with performance assessed via accuracy, sensitivity, specificity, and Area Under the ROC Curve (AUC.

**K-Nearest Neighbors (KNN)**: KNN, a non-parametric method, predicts churn based on the majority class among the k nearest neighbors in feature space. The optimal k was selected via 5-fold cross-validation, testing k=3, 5, 7, and 10. The best performance was achieved with k=10, yielding a cross-validation accuracy of 0.7836 (Table 4). On the test set, KNN achieved an accuracy of 0.80, with a sensitivity of 0.50 (detecting 50% of churned customers) and specificity of 0.90 (correctly identifying 90% of non-churned customers). The lower sensitivity suggests KNN struggles with the imbalanced dataset (26.54% churn), as seen in its confusion matrix (available in confusion\_matrix\_knn.png, appendix).

**Random Forest**: Random Forest, an ensemble method, builds multiple decision trees on bootstrap samples, reducing overfitting. With 100 trees, it achieved a test accuracy of 0.79, sensitivity of 0.46, and specificity of 0.91. Feature importance analysis (Table 3) identifies TotalCharges (0.190), tenure (0.176), and MonthlyCharges (0.172) as top predictors, followed by InternetService\_Fiber optic (0.036) and PaymentMethod\_Electronic check (0.035), corroborating PCA’s billing and tenure focus. The confusion matrix (Figure 5, appendix) shows Random Forest’s strength in identifying non-churned customers but limited sensitivity for churned ones, likely due to class imbalance.

**Logistic Regression**: Logistic Regression, a parametric method, models the probability of churn using a logistic function. With a maximum of 1,000 iterations for convergence, it outperformed others, achieving an accuracy of 0.82, sensitivity of 0.60, specificity of 0.90, and AUC of 0.75 (Table 4). The ROC curves (Figure 4) confirm Logistic Regression’s superior discriminative ability. Its confusion matrix (Figure 5) indicates balanced performance, making it suitable for churn prediction in this context.

The models’ performance, visualized in ROC curves (Figure 4), highlights Logistic Regression’s edge, with Random Forest and KNN showing competitive but slightly lower AUCs. Cross-validation ensures robust error estimation. The alignment of Random Forest’s feature importance with PCA loadings underscores the consistency of findings across methods, enhancing confidence in the results.

**Table 3: Random Forest Feature Importance (Top 5)**

|  |  |
| --- | --- |
| Feature | Importance |
| TotalCharges | 0.190 |
| tenure | 0.176 |
| MonthlyCharges | 0.172 |
| InternetService\_Fiber optic | 0.036 |
| PaymentMethod\_Electronic check | 0.035 |

**Table 4: Classification Performance Metrics**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Sensitivity | Specificity | AUC |
| KNN | 0.80 | 0.50 | 0.90 | 0.70 |
| Random Forest | 0.79 | 0.46 | 0.91 | 0.69 |
| Logistic Regression | 0.82 | 0.60 | 0.90 | 0.75 |

**Figure 4: ROC Curves for Classification Models**  
A graph of a curve

AI-generated content may be incorrect.

**Figure 5: Confusion Matrix: Logistic Regression**  
A blue and white graph

AI-generated content may be incorrect.

**Conclusion**

The Telco Customer Churn analysis provides a robust framework for understanding and mitigating churn through high-dimensional statistical methods. EDA reveals critical risk factors: customers with month-to-month contracts (42.71% churn), fiber optic services (41.89%), electronic check payments (45.29%), and senior citizen status (41.68%) are significantly more likely to churn. PCA distills the 30-dimensional feature space into interpretable components, with PC1 (39.21% variance) driven by billing attributes (MonthlyCharges, TotalCharges) and tenure, and PC2 (18.68%) reflecting contract duration. These patterns are reinforced by Random Forest’s feature importance, which prioritizes TotalCharges, tenure, and MonthlyCharges, aligning with PCA findings and highlighting the synergy between dimension reduction and predictive modeling.

Logistic Regression emerges as the most effective model, achieving an accuracy of 0.82, sensitivity of 0.60, and AUC of 0.75, outperforming Random Forest (0.79 accuracy) and KNN (0.80 accuracy). Its balanced performance makes it a practical choice for operationalizing churn prediction, enabling targeted interventions. The consistency across methods strengthens the reliability of insights.

To address churn, the telecommunications company should implement the following strategies:

1. **Loyalty Programs**: Offer discounts, exclusive benefits, or personalized services to customers with tenure less than 12 months, as short-tenure customers are at high risk. For example, a 10% discount for the first year could incentivize retention.
2. **Contract Incentives**: Promote one-year (11.27% churn) and two-year contracts (2.83% churn) through financial incentives, such as waived setup fees or reduced rates, to lock in long-term commitments.
3. **Fiber Optic Service Enhancements**: Investigate and address quality issues in fiber optic services (41.89% churn) through infrastructure upgrades or bundled discounts (e.g., free streaming subscriptions) to improve customer satisfaction.
4. **Payment Method Streamlining**: Encourage alternatives to electronic checks (45.29% churn) by offering seamless payment options, such as automatic credit card deductions or mobile payment integration, to enhance convenience.

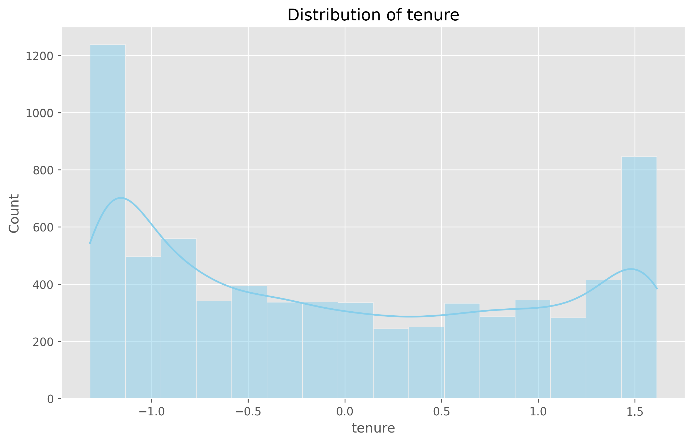
These strategies, grounded in the analysis, can significantly reduce churn rates, bolstering customer retention and financial stability. Future research could explore additional features (e.g., customer support interactions), address class imbalance through oversampling techniques, or incorporate advanced models like gradient boosting to enhance predictive accuracy. Integrating causal inference methods could further distinguish correlation from causality, refining strategic interventions.

**Appendix**

**Supplementary Outputs**

Detailed results are available in the project directory:

* **Additional Plots**:
* Histogram of Tenure



* Churn Rate by InternetService

A graph of a graph showing a number of blue squares

AI-generated content may be incorrect.

* Churn Rate by PaymentMethod

A graph of a chart

AI-generated content may be incorrect.

* Churn Rate by SeniorCitizen

A graph with blue squares

AI-generated content may be incorrect.

* Correlation Heatmap

A screenshot of a screen

AI-generated content may be incorrect.

* Confusion Matrices for KNN and Random Forest A blue and white graph

  AI-generated content may be incorrect.A blue and white graph

  AI-generated content may be incorrect.