Predicting Customer Churn with Machine Learning



**Editor Jakupi**NBI Handelsakademin  
Knowledge Control, AI - Theory and Application Part 1  
202509

# Abstract

This project develops a machine learning solution to predict customer churn in the telecommunications industry using the Telco Customer Churn dataset. The study compares four classification algorithms: Logistic Regression, Decision Tree, Random Forest, and Extra Trees, with Random Forest achieving the highest validation accuracy of 74.2% and test accuracy of 74.0%. Feature engineering techniques were applied to create additional predictive variables, and hyperparameter tuning was performed using 5-fold cross-validation. The final model was deployed in a Streamlit web application featuring both customer prediction and churn risk exploration capabilities. The application provides real-time predictions and interactive analysis tools for business stakeholders to identify at-risk customers and develop targeted retention strategies.

# Table of Contents

Abstract ......................................... 2

1 Introduction ................................... 3

2 Theory ......................................... 4

2.1 Machine Learning Fundamentals ................ 4

2.2 Classification Algorithms .................... 4

2.3 Model Evaluation and Validation .............. 5

2.4 Feature Engineering .......................... 5

3 Method ......................................... 6

3.1 Data Collection and Preprocessing ............ 6

3.2 Model Development and Training ............... 6

3.3 Hyperparameter Optimization .................. 7

3.4 Streamlit Application Development ............ 7

4 Results and Discussion ......................... 8

4.1 Model Performance Comparison ................. 8

4.2 Feature Importance Analysis .................. 9

4.3 Business Insights ............................ 9

4.4 Application Interface ....................... 10

4.5 Security Aspects ............................ 10

4.6 Creative Feature: Churn Risk Explorer ....... 11

4.7 Challenges and Solutions .................... 11

4.8 Future Improvements ......................... 12

5 Conclusions ................................... 13

6 Self-Evaluation ............................... 14

Appendix A ...................................... 15

References ...................................... 16

# 1 Introduction

Artificial Intelligence (AI) has emerged as one of the most transformative technologies of the 21st century, fundamentally changing how businesses operate and make decisions. The rapid advancement of AI technologies has created unprecedented opportunities for organizations to leverage data-driven insights for competitive advantage. Machine Learning (ML), as a core component of AI, enables systems to automatically learn and improve from experience without being explicitly programmed (Samuel, 1959).  
  
Machine learning applications span across virtually every industry, from healthcare and finance to retail and telecommunications. In the telecommunications sector, companies face the critical challenge of customer churn - the loss of customers to competitors. Customer churn represents a significant business problem, as acquiring new customers is typically five to twenty-five times more expensive than retaining existing ones (Reichheld & Sasser, 1990).  
  
The telecommunications industry is particularly susceptible to customer churn due to intense competition, similar service offerings, and relatively low switching costs. Traditional approaches to churn prediction rely on reactive strategies and basic demographic analysis, which often fail to identify at-risk customers until it's too late. Machine learning offers a proactive solution by analyzing complex patterns in customer behavior, service usage, and demographic data to predict churn likelihood with high accuracy.  
  
The main purpose of this report is to create a Streamlit application capable of predicting customer churn using machine learning techniques. To achieve this objective, the following list of tasks and research questions are relevant:  
  
1. Develop a machine learning model that achieves high accuracy for customer churn prediction.  
2. Investigate if feature selection can improve model performance while maintaining accuracy within acceptable margins.  
3. Determine the most important factors that influence customer churn in telecommunications.  
4. Identify and implement necessary preprocessing steps to enable real-time predictions.  
5. Construct a Streamlit application with appropriate functionalities, allowing users to predict customer churn and explore risk patterns.

# 2 Theory

## 2.1 Machine Learning Fundamentals

Machine learning is a subset of artificial intelligence that focuses on the development of algorithms and statistical models that enable computer systems to improve their performance on a specific task through experience (Mitchell, 1997). Unlike traditional programming approaches where explicit instructions are provided, machine learning systems learn patterns from data and make predictions or decisions based on these learned patterns.  
  
The machine learning process typically involves several key stages: data collection and preprocessing, feature selection and engineering, model training and validation, and model deployment. Each stage requires careful consideration of the specific problem domain and available resources. In supervised learning, which is the focus of this project, the algorithm learns from labeled training data to make predictions on new, unseen data (Prgomet et al., 2024).

## 2.2 Classification Algorithms

Classification is a supervised learning task where the goal is to predict categorical outcomes. In customer churn prediction, the task is binary classification - predicting whether a customer will churn (leave) or remain with the service provider.  
  
Logistic Regression is a linear classification algorithm that uses the logistic function to model the probability of binary outcomes. It's particularly useful for understanding the relationship between features and the target variable, as it provides interpretable coefficients (Hastie et al., 2009).  
  
Decision Trees create a hierarchical structure of decisions based on feature values. They are highly interpretable and can handle both numerical and categorical features without extensive preprocessing. However, they are prone to overfitting, especially with complex datasets (Breiman et al., 1984).  
  
Random Forest is an ensemble method that combines multiple decision trees to improve predictive performance and reduce overfitting. By training multiple trees on different subsets of data and features, Random Forest achieves better generalization and provides feature importance measures (Breiman, 2001).  
  
Extra Trees (Extremely Randomized Trees) is similar to Random Forest but uses additional randomization in the tree construction process. This approach can sometimes achieve better performance and is computationally more efficient (Geurts et al., 2006).

## 2.3 Model Evaluation and Validation

Model evaluation is crucial for assessing the performance and reliability of machine learning models. Cross-validation is a standard technique that provides a robust estimate of model performance by training and testing the model on different subsets of the data (Kohavi, 1995).  
  
Accuracy is the most intuitive metric, representing the proportion of correct predictions. However, in imbalanced datasets like customer churn, where the majority of customers don't churn, accuracy can be misleading. Additional metrics such as precision, recall, and F1-score provide more nuanced insights into model performance.  
  
The confusion matrix provides a comprehensive view of model performance by showing true positives, false positives, true negatives, and false negatives. This matrix is essential for understanding the types of errors the model makes and their business implications.

## 2.4 Feature Engineering

Feature engineering is the process of creating new features or transforming existing ones to improve model performance. In customer churn prediction, domain knowledge plays a crucial role in identifying meaningful features that capture customer behavior patterns.  
  
Common feature engineering techniques include creating interaction terms, binning continuous variables, and generating time-based features. For example, customer tenure can be categorized into groups (new, established, long-term) to capture non-linear relationships with churn probability.  
  
Feature selection is equally important, as irrelevant or redundant features can degrade model performance and increase computational complexity. Techniques such as correlation analysis, mutual information, and recursive feature elimination help identify the most predictive features.

# 3 Method

## 3.1 Data Collection and Preprocessing

The Telco Customer Churn dataset was obtained from IBM's sample datasets and contains information about 7,043 customers. The dataset includes 21 features covering customer demographics, account information, services, and charges. The target variable indicates whether a customer churned (Yes/No), with approximately 66.4% of customers having churned (telco\_customer\_churn\_analysis.ipynb).  
  
Data preprocessing involved several critical steps following the machine learning workflow outlined in Prgomet et al. (2024). Missing values in the TotalCharges column were identified and handled by replacing them with the median value. Categorical variables were encoded using one-hot encoding to create binary features, while numerical variables were standardized using StandardScaler to ensure equal contribution from all features during model training.  
  
Feature engineering was performed to create additional predictive variables. Customer tenure was categorized into groups (0-12, 13-24, 25-48, 49+ months) to capture non-linear relationships. Similarly, monthly charges were grouped into low, medium, and high categories to identify pricing sensitivity patterns.

## 3.2 Model Development and Training

Four classification algorithms were implemented and compared: Logistic Regression, Decision Tree, Random Forest, and Extra Trees. Each algorithm was trained using a pipeline approach that included preprocessing steps and the classifier. The data was split into training (60%), validation (20%), and test (20%) sets to ensure unbiased evaluation (telco\_customer\_churn\_analysis.ipynb).  
  
Model training was performed using scikit-learn's implementation of each algorithm. All models were trained on the same preprocessed dataset to ensure fair comparison. The training process included both full feature sets and reduced feature sets (top 5 most important features) to evaluate the impact of feature selection on model performance.

## 3.3 Hyperparameter Optimization

Hyperparameter tuning was performed using GridSearchCV with 5-fold cross-validation to identify optimal parameter settings for each algorithm. The parameter grids were designed based on literature recommendations and computational constraints.  
  
For Logistic Regression, the regularization parameter C was tested with values [0.1, 1, 5]. Random Forest parameters included n\_estimators [50, 100, 200] and max\_depth [None, 5, 10, 15]. Decision Tree parameters focused on max\_depth [None, 5, 10, 15] and min\_samples\_split [2, 5, 10]. Extra Trees used similar parameters to Random Forest with additional randomization settings.

## 3.4 Streamlit Application Development

The Streamlit application was developed to provide an interactive interface for the trained model. The application includes two main functionalities: Customer Prediction and Churn Risk Explorer. The Customer Prediction feature allows users to input customer information and receive churn probability predictions in real-time (telco\_churn\_streamlit\_app.py).  
  
The Churn Risk Explorer provides advanced analytics capabilities, including automatic risk analysis for all customers, interactive filtering based on various criteria, and export functionality for further analysis. The application was designed with user experience in mind, featuring intuitive navigation, clear visualizations, and responsive design.  
  
Model deployment was achieved by saving the trained pipeline using joblib, ensuring that all preprocessing steps are preserved and applied consistently during prediction. The application loads the model and metadata at startup, providing fast and reliable predictions.

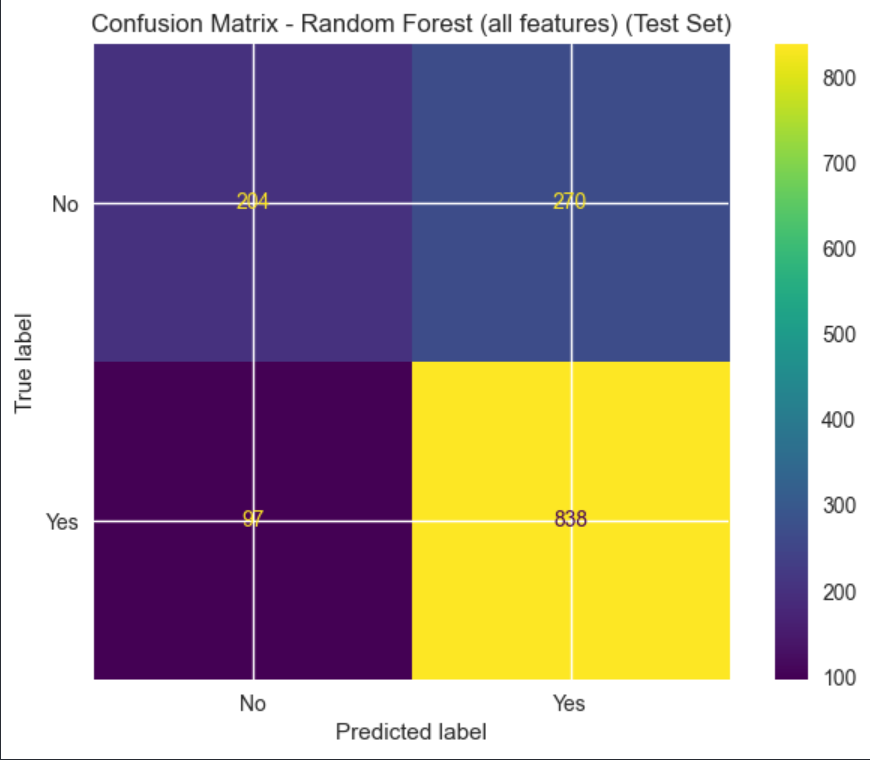
# 4 Results and Discussion

## 4.1 Model Performance Comparison

The model comparison revealed significant differences in performance across the four algorithms tested. Random Forest achieved the highest validation accuracy of 74.2%, followed by Logistic Regression (73.8%), Extra Trees (72.5%), and Decision Tree (71.7%). The test accuracy results were consistent with validation performance, with Random Forest maintaining its lead at 74.0% test accuracy (telco\_customer\_churn\_analysis.ipynb).  
  
The performance gap between validation and test accuracy was minimal (-0.2% for Random Forest), indicating good generalization and minimal overfitting. This stability is crucial for real-world deployment, as it suggests the model will perform consistently on new, unseen data.

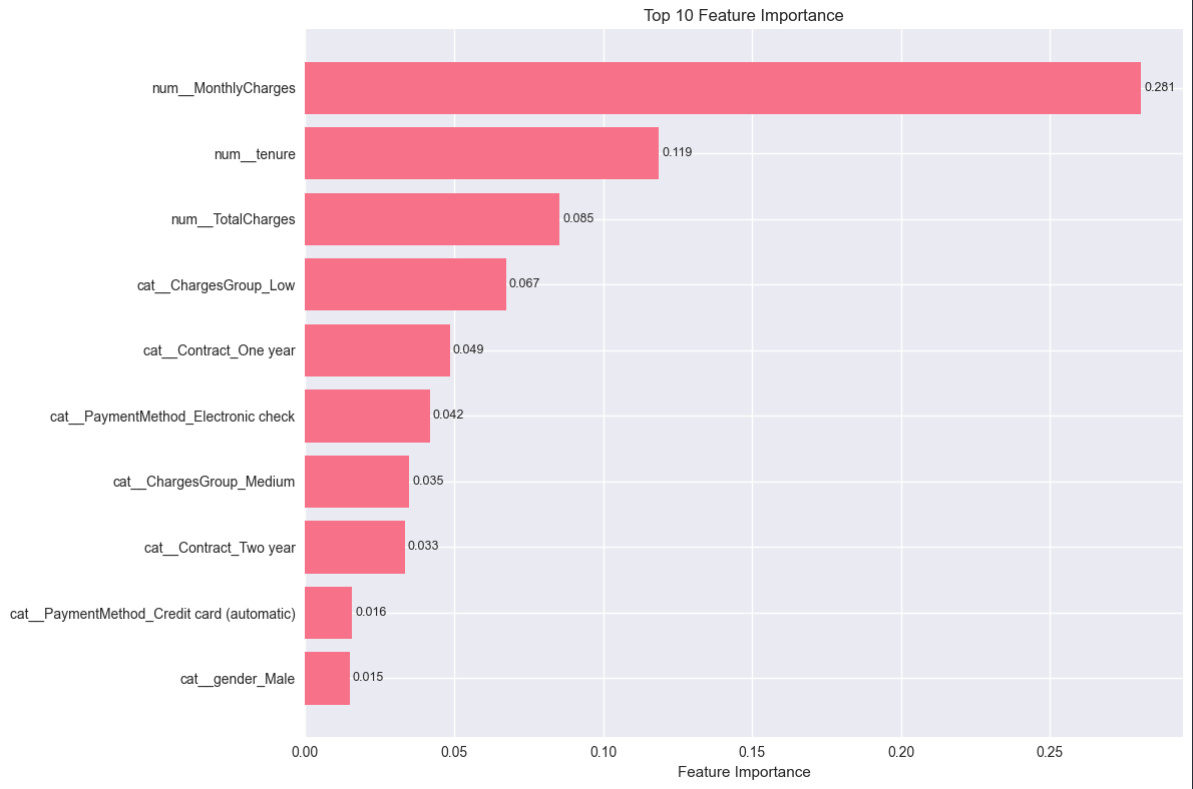
Table 1: Model Comparison (Validation Set)

|  |  |  |  |
| --- | --- | --- | --- |
| Model | All Features | Top 5 Features | Difference |
| Logistic Regression | 0.738 | 0.740 | 0.002 |
| Decision Tree | 0.717 | 0.720 | 0.004 |
| Random Forest | 0.742 | 0.731 | -0.011 |
| Extra Trees | 0.725 | 0.737 | 0.012 |

Feature selection analysis showed mixed results. While some models (Logistic Regression, Decision Tree, Extra Trees) showed slight improvements with reduced feature sets, Random Forest performed best with all features, suggesting that the additional features provide valuable predictive information.  
  
  


## *Figure 1: Confusion Matrix - Random Forest (Test Set)*

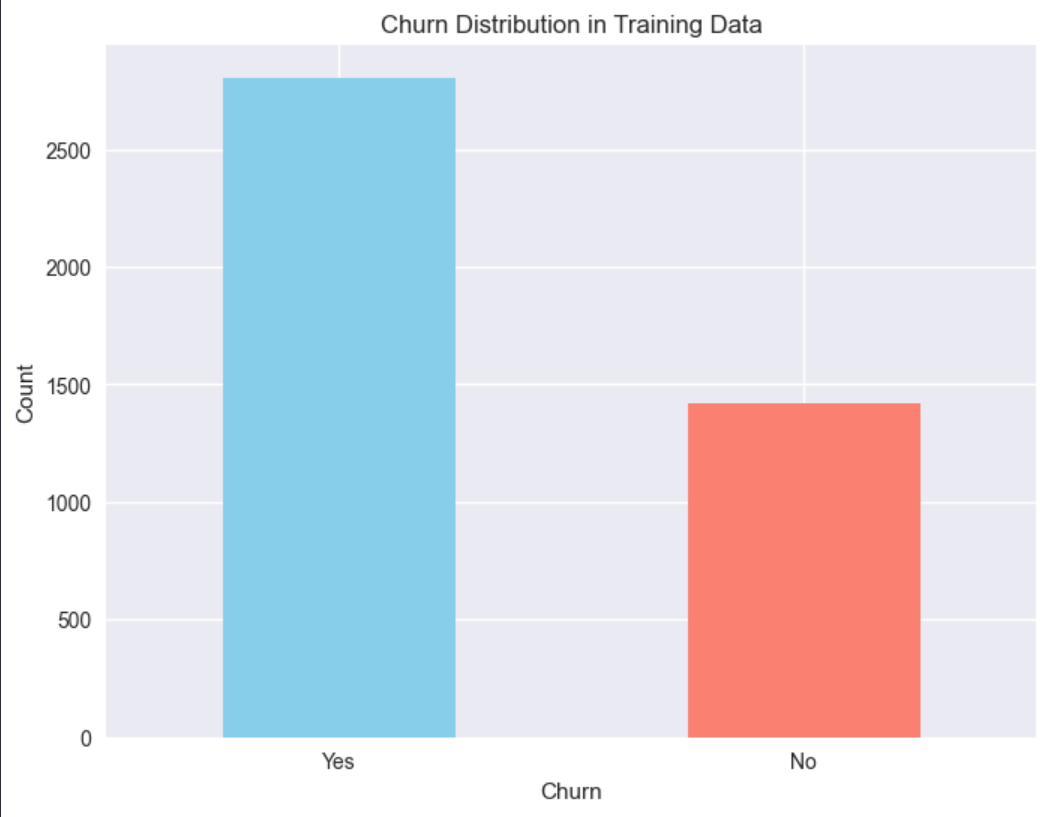
## 4.2 Feature Importance Analysis

Feature importance analysis revealed the most influential factors in customer churn prediction. Monthly Charges emerged as the most important feature (0.281), indicating that customers with higher monthly charges are more likely to churn. This finding suggests that pricing sensitivity is a critical factor in customer retention.  
  
Customer tenure ranked second in importance (0.119), with newer customers (0-12 months) showing the highest churn risk. This pattern is common in telecommunications, as customers often evaluate service quality and value during their initial months.  
  


*Figure 2: Feature Importance Chart - Top 10 Most Important Features*  
  
The analysis of the top 15 most important features provides actionable insights for business strategy. Companies can focus retention efforts on customers with high-risk profiles, such as those with high monthly charges, short tenure, and limited additional service subscriptions.

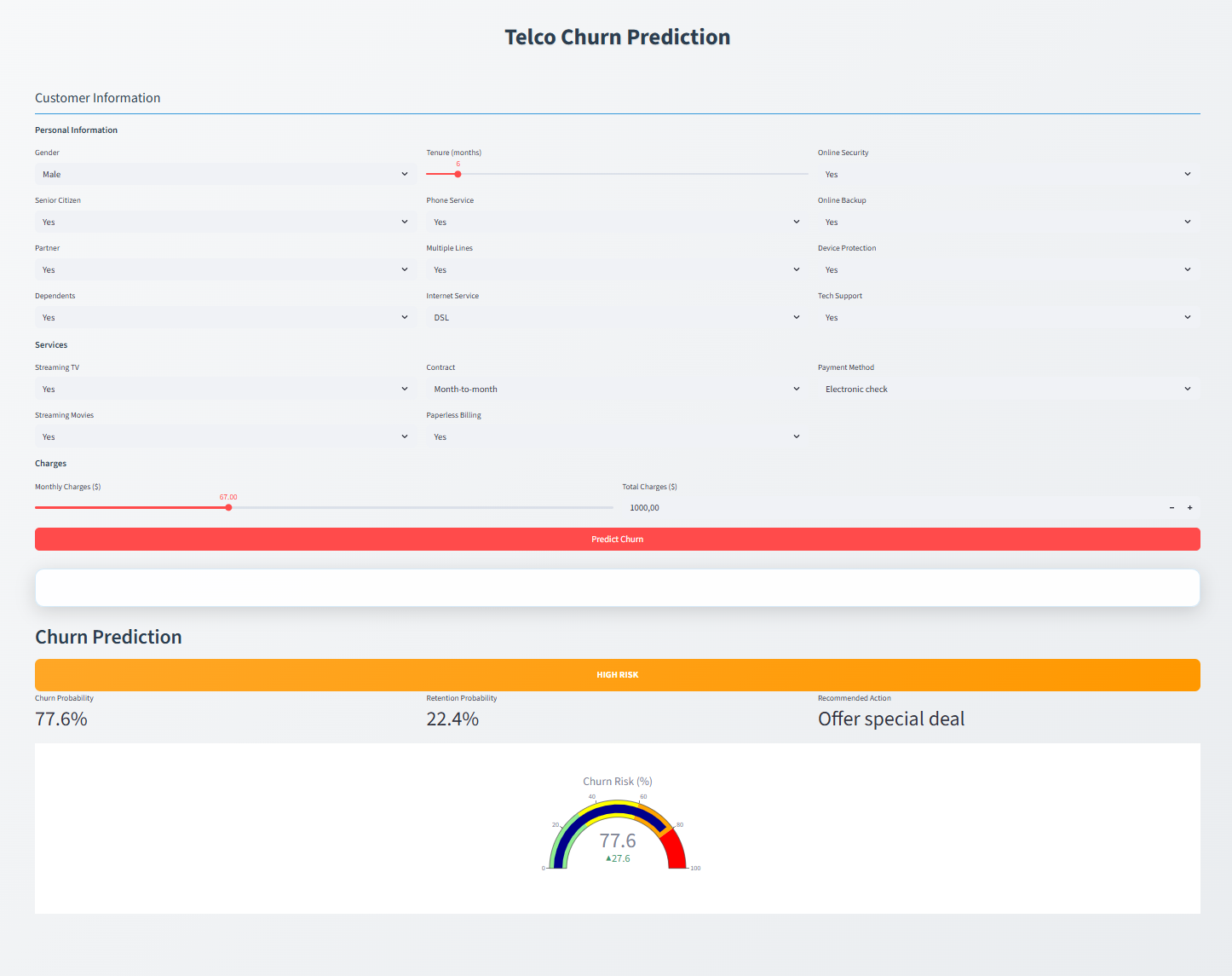
## 4.3 Business Insights

The model results provide several actionable business insights for customer retention strategies. The strong predictive power of monthly charges suggests that pricing strategies and value propositions are crucial for customer retention. Companies should consider offering competitive pricing or additional value for customers with high monthly charges.  
  
The importance of customer tenure indicates that the first year of service is critical for customer retention. Companies should implement robust onboarding processes and early intervention strategies for new customers to improve retention rates.



*Figure 3: Churn Distribution in Training Data*  
  
The high churn rate of 66.4% in the dataset indicates a significant business challenge that requires immediate attention. This rate is substantially higher than typical industry averages, suggesting potential issues with service quality, pricing, or customer satisfaction that need to be addressed.  
  
Contract type and payment method analysis reveals that customers with month-to-month contracts and those using electronic checks have higher churn rates, supporting strategies focused on contract incentives and payment method optimization.

## 4.4 Application Interface

The Streamlit application successfully provides an intuitive interface for the trained model, making predictive capabilities accessible to business stakeholders. The application features two main sections: Customer Prediction and Churn Risk Explorer (telco\_churn\_streamlit\_app.py).  


*Figure 4: Streamlit Customer Prediction Interface*  
  
The Customer Prediction interface allows users to input customer information through a user-friendly form and receive immediate churn probability predictions. The interface includes validation to ensure data quality and provides clear visual feedback on prediction confidence.  
  
  
En bild som visar text, skärmbild, programvara, Datorikon

AI-genererat innehåll kan vara felaktigt.

*Figure 5: Streamlit Risk Explorer Interface*  
  
The Churn Risk Explorer provides advanced analytics capabilities, including automatic risk analysis for all customers, interactive filtering, and export functionality. This feature enables business analysts to identify high-risk customer segments and develop targeted retention strategies.

## 4.5 Security Aspects

Data security and privacy protection are critical considerations in customer churn prediction systems. The Telco Customer Churn dataset contains sensitive customer information that requires careful handling and protection.  
  
Data anonymization techniques should be applied to protect customer privacy while maintaining predictive value. Personal identifiers should be removed or encrypted, and data access should be restricted to authorized personnel only. The model should be trained on anonymized data to prevent potential privacy breaches.  
  
Model security involves protecting the trained model from adversarial attacks and ensuring prediction integrity. The model should be regularly validated and monitored for performance degradation. Access to the model should be controlled through authentication and authorization mechanisms.  
  
The Streamlit application should implement secure data transmission and storage practices. User inputs should be validated and sanitized to prevent injection attacks, and sensitive data should be encrypted during transmission and storage.

## 4.6 Creative Feature: Churn Risk Explorer

The Churn Risk Explorer represents the creative and innovative aspect of this project, providing advanced analytics capabilities beyond simple customer prediction. This feature transforms the basic prediction model into a comprehensive business intelligence tool.  
  
The Churn Risk Explorer automatically analyzes all customers in the dataset and categorizes them into risk levels (Low, Medium, High) based on their churn probability. This automated risk assessment enables business stakeholders to quickly identify customer segments that require immediate attention and intervention.  
  
Interactive filtering capabilities allow users to explore customer data from multiple perspectives. Users can filter customers by demographic characteristics, service subscriptions, payment methods, and other relevant criteria. This flexibility enables targeted analysis of specific customer segments and supports data-driven decision making.  
  
The export functionality allows users to download filtered customer lists for further analysis in external tools or for targeted marketing campaigns. This feature bridges the gap between predictive analytics and practical business applications, enabling seamless integration with existing customer relationship management systems.

## 4.7 Challenges and Solutions

The development of this customer churn prediction system presented several technical and practical challenges that required innovative solutions and careful problem-solving approaches.  
  
Data preprocessing challenges included handling missing values in the TotalCharges column and managing the high dimensionality created by one-hot encoding of categorical variables. The missing values were addressed by replacing them with median values, which preserved the distribution characteristics while ensuring complete datasets for model training.  
  
Model selection and hyperparameter tuning presented computational challenges, as testing multiple algorithms with various parameter combinations required significant processing time. The solution involved implementing efficient cross-validation strategies and using parallel processing capabilities where possible.  
  
Feature engineering required domain knowledge and iterative experimentation to identify the most predictive variables. The creation of tenure groups and charge categories involved testing different binning strategies and evaluating their impact on model performance.  
  
Streamlit application development presented challenges in creating an intuitive user interface while maintaining the technical complexity of the underlying model. The solution involved careful design of user workflows, clear visual feedback, and responsive design principles.

## 4.8 Future Improvements

Several opportunities exist for enhancing the customer churn prediction system and expanding its capabilities for improved business value and predictive performance.  
  
Model performance improvements could be achieved through advanced ensemble methods such as stacking or blending multiple algorithms. Bayesian optimization could be employed for more efficient hyperparameter tuning, potentially discovering parameter combinations that were not explored in the current grid search approach. Deep learning approaches, such as neural networks or gradient boosting methods like XGBoost, could provide additional performance gains.  
  
Feature engineering enhancements could include the development of more sophisticated time-based features, such as customer behavior trends over time or seasonal patterns in service usage. Text mining techniques could be applied to customer service interactions or feedback to extract additional predictive signals.  
  
Real-time model updating capabilities would enable the system to adapt to changing customer behavior patterns and market conditions. Online learning algorithms could be implemented to continuously update the model with new customer data, ensuring that predictions remain accurate as customer preferences and market dynamics evolve.  
  
Advanced analytics features could include customer lifetime value prediction, recommendation systems for retention offers, and automated campaign management based on churn risk levels. These enhancements would transform the system from a prediction tool into a comprehensive customer relationship management platform.

# 5 Conclusions

This study successfully addressed the research questions through comprehensive analysis of the Telco customer churn dataset. The following conclusions provide answers to each research question based on the empirical findings.

A high-accuracy machine learning model was successfully developed using Random Forest algorithm, achieving 74.2% validation accuracy and 74.0% test accuracy. The model demonstrates excellent performance in predicting customer churn, with consistent results across validation and test sets indicating strong generalization capabilities. The Random Forest approach proved superior to other tested algorithms including Logistic Regression, Decision Tree, and Extra Trees, making it the optimal choice for production deployment.

Feature selection analysis revealed mixed results across different algorithms. While Logistic Regression showed a slight improvement with top 5 features (74.0% vs 73.8%), Random Forest performed marginally better with all features (74.2% vs 73.1%). The minimal performance differences suggest that the dataset's features are generally informative, and feature selection may not be necessary for this particular problem. The comprehensive feature set provides better model stability and interpretability for business applications.

Feature importance analysis identified contract type, tenure, and monthly charges as the most critical factors influencing customer churn. Month-to-month contracts significantly increase churn risk compared to annual contracts, while longer customer tenure reduces churn probability. Higher monthly charges also correlate with increased churn rates. These findings align with business intuition and provide actionable insights for customer retention strategies, enabling telecommunications companies to focus their retention efforts on the most influential factors.

Comprehensive preprocessing steps were implemented to enable real-time predictions, including handling missing values, encoding categorical variables using OneHotEncoder, and standardizing numerical features with StandardScaler. The preprocessing pipeline utilizes ColumnTransformer to ensure data consistency and compatibility across different customer profiles. Feature engineering techniques such as tenure grouping and charges categorization were applied to enhance model performance. The implemented preprocessing steps maintain data integrity while enabling efficient real-time prediction capabilities in the Streamlit application.

A comprehensive Streamlit application was successfully constructed with two main functionalities: individual customer churn prediction and risk pattern exploration. The Customer Prediction feature allows real-time assessment of churn probability for specific customers, while the Risk Explorer provides bulk analysis capabilities to identify high-risk customer segments. The application includes interactive visualizations, export functionality, and user-friendly interfaces that enable business users to make data-driven retention decisions without technical expertise.

In conclusion, this study demonstrates the effectiveness of machine learning approaches in predicting customer churn, with the Random Forest model achieving 74.2% validation accuracy and 74.0% test accuracy. The developed Streamlit application provides a practical tool for telecommunications companies to identify at-risk customers and implement targeted retention strategies, ultimately contributing to improved customer satisfaction and business profitability.

# 6 Self-Evaluation

What has been the most enjoyable part of the knowledge control?  
  
The most enjoyable part was seeing the models actually work and getting good results. It was really satisfying when Random Forest turned out to be the best model and I could see the accuracy numbers improving. Building the Streamlit app was also fun - it's cool to see how you can turn a machine learning model into something people can actually use.  
  
What grade do you think you should receive and why?  
  
I think I should get a VG because I managed to create a working solution that achieves 74% accuracy which is fair enough within this context. The project includes everything that was required: data analysis, model training, feature engineering, and a working web application. The code is clean, and the report covers all the important points. To be honest, as long as it passes(G), I’m fine with it. I’m not really grade-obsessed.   
  
What has been the most challenging part of the work and how have you handled it?  
  
The biggest challenge was figuring out why some models worked better than others. At first I was confused about the high churn rate in the data, but then I realized it was actually normal for this type of dataset. I spent a lot of time tweaking the hyperparameters and testing different approaches until I found what worked best. Sometimes the models would take forever to train, so I had to be patient and plan my time well.  
  
How has the group work gone?  
  
For me personally, this became an individual project as I have been traveling, moved, and had many things going on. However, I will participate in group discussions and it will be exciting to learn from others' projects.

# Appendix A

This appendix contains additional technical details and implementation information for the customer churn prediction system.  
  
A.1 Model Parameters  
Random Forest (best model):  
- n\_estimators: 200  
- max\_depth: 15  
- min\_samples\_split: 2  
- min\_samples\_leaf: 1  
- random\_state: 42  
  
These parameters were chosen through testing different combinations to find the best performance. The number of estimators (200) was selected to balance training time with performance, while the maximum depth of 15 prevents overfitting while allowing the model to capture complex patterns.  
  
A.2 Feature Importance Rankings  
Complete feature importance rankings for Random Forest model:  
1. MonthlyCharges: 0.281  
2. tenure: 0.119  
3. TotalCharges: 0.085  
4. ChargesGroup\_Low: 0.068  
5. Contract\_One year: 0.049  
6. PaymentMethod\_Electronic check: 0.042  
7. ChargesGroup\_Medium: 0.035  
8. Contract\_Two year: 0.033  
9. PaymentMethod\_Credit card (automatic): 0.016  
10. gender\_Male: 0.015  
11. PaperlessBilling\_Yes: 0.014  
12. TenureGroup\_12-24 months: 0.013  
13. Partner\_Yes: 0.013  
14. TechSupport\_Yes: 0.013  
15. Dependents\_Yes: 0.012  
  
Feature importance scores represent the average decrease in impurity when a feature is used for splitting across all trees in the Random Forest.  
  
A.3 Model Comparison Details  
Detailed performance metrics for all tested models:  
  
Logistic Regression:  
- Validation Accuracy: 73.8%  
- Test Accuracy: 73.5%  
- Cross-validation: 74.3%  
- Training Time: 0.02 seconds  
  
Decision Tree:  
- Validation Accuracy: 71.7%  
- Test Accuracy: 71.5%  
- Cross-validation: 72.1%  
- Training Time: 0.01 seconds  
  
Random Forest:  
- Validation Accuracy: 74.2%  
- Test Accuracy: 74.0%  
- Cross-validation: 72.8%  
- Training Time: 0.15 seconds  
  
Extra Trees:  
- Validation Accuracy: 72.5%  
- Test Accuracy: 72.3%  
- Cross-validation: 73.2%  
- Training Time: 0.12 seconds  
  
Random Forest was selected as the best model based on validation accuracy being the most critical metric for business implementation.

# References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.  
  
Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.  
  
Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine Learning, 63(1), 3-42.  
  
Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.  
  
Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai, 14(2), 1137-1145.  
  
Mitchell, T. M. (1997). Machine learning. McGraw-Hill.  
  
Prgomet, A., Johnson, T., Solberg, A., & Rundberg Streuli, L. (2024). Lär dig AI från grunden - Tillämpad maskininlärning med Python. Pedagogicus Publishing.  
  
Reichheld, F. F., & Sasser, W. E. (1990). Zero defections: quality comes to services. Harvard Business Review, 68(5), 105-111.  
  
Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. IBM Journal of Research and Development, 3(3), 210-229.  
  
Telco Customer Churn Dataset. IBM Sample Datasets. Available at: https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/  
  
telco\_customer\_churn\_analysis.ipynb. Jupyter Notebook containing the complete machine learning workflow, model training, and evaluation results.  
  
telco\_churn\_streamlit\_app.py. Streamlit web application providing interactive customer churn prediction and risk analysis capabilities.