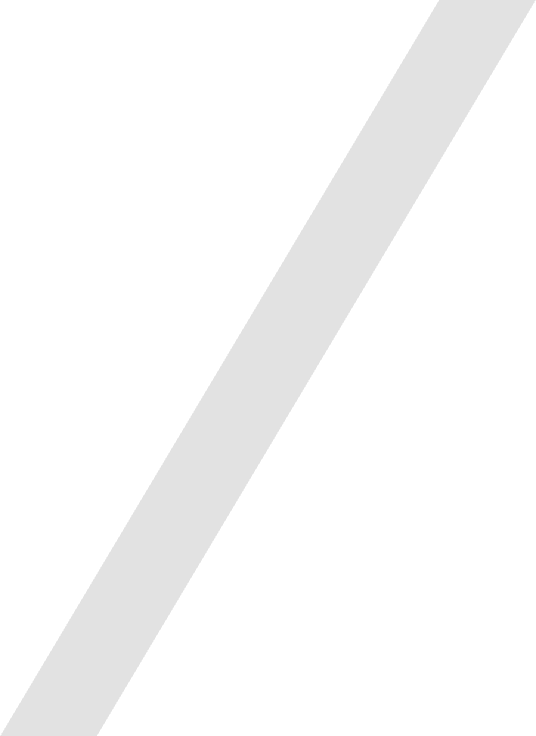
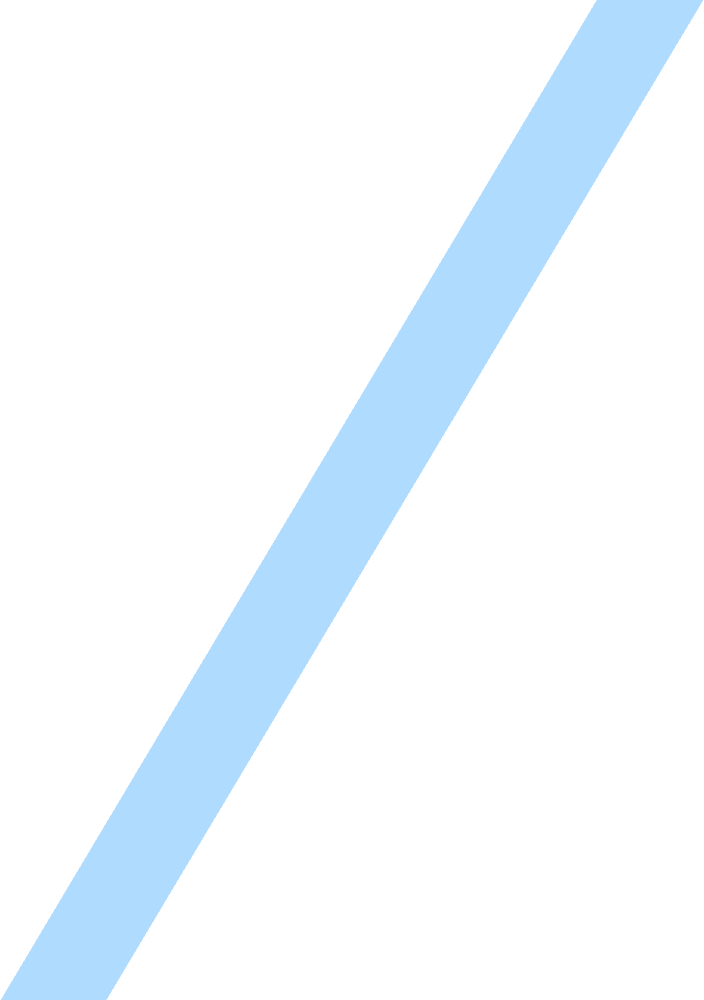
|  |
| --- |
| TECHNICAL REPORT |

|  |
| --- |
| Electrical & Computer Engineering & Computer Science (ECECS) |

|  |  |
| --- | --- |
| Fall 23 |  |



Contents

[Customer Churn Prediction for a Telecom Company 2](#_Toc96341549)

[Executive Summary 2](#_Toc96341550)

[Technical Report 3](#_Toc96341551)

[Highlights of Project 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341552)

[Submitted on: 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341553)

[Abstract 4](#_Toc96341554)

[Methodology 5](#_Toc96341555)

[Results Section](#_Toc96341556) 8

[Discussion](#_Toc96341557) 10

[Conclusion](#_Toc96341558) 10

[Contributions/References](#_Toc96341559) 11

|  |
| --- |
| Heart Disease Prediction Using Logistic Regression |

|  |  |  |
| --- | --- | --- |
| Executive Summary This project leverages existing machine learning models to process telecom customer data and predict the likelihood of churn. The proposed system entails a comprehensive approach, starting with a preprocessing stage that involves data segmentation, eliminating irrelevant information, and mitigating potential biases. Following this, a classification model is trained using transfer learning techniques to enhance predictive accuracy. The final stage involves a thorough analysis and interpretation of the results, providing actionable insights to telecom providers for effective churn management and customer retention strategies. | | |
| person at a table writing in a notebook with people around | | |
| **Team Members:**  **Giri M**  **Sreehar S**  **Venkata Susmitha**  **Sowjanya k** |  | **Questions?**  Contact : +(203) 843-3889 |

|  |
| --- |
|  |

|  |  |
| --- | --- |
| * Strategic Tool Selection: Leveraged Python's libraries, Git/GitHub for collaboration, and PythonAnywhere for deployment, ensuring a comprehensive software approach. * Empowering Telecom Strategies: Enabled proactive churn identification, fostering targeted retention and operational optimization in the dynamic telecom sector. * User-Friendly Web App: Created an intuitive Flask and HTML/CSS-based interface, enabling seamless interaction and valuable insights. * Dataset includes demographic data and factors (2000+ records, 667 rows and 20 features) * Model shows 90% accuracy towards churn. |  |
|  |



# Technical Report

***Customer Churn Prediction for a Telecom Company***

## 

Submitted on: 12/06/2023

## Abstract

This technical report delves into a robust customer churn prediction system for a telecom company. Here, getting a new customer base is costlier than holding the current customers where churn is the process of customers switching from one firm to another in a given stipulated time. The major goal of this system is to analyze the diversified machine learning algorithms which are required to develop customer churn prediction models and identify churn reasons to give them with retention strategies and plans. In this system, leave subscriptions collects customers' data by applying classification algorithms such as Random Forest (RF), machine learning techniques such as decision tree Classifier. It offers an efficient business model that analyzes customer churn data and gives accurate predictions of churn customers so that business management may act within the churn period to stop churn as well as loss in profit. System achieves an accuracy of 99 % using the random forest classifier for churn predicts the classifier matrix has achieved a precision of 99 % with a recall factor of 99 % along with received overall accuracy of 99.09 %. Likewise, our research work improves churn prediction, scope other business fields, and provide prediction models to hold their existing customers customer service and avoid churn effectively.

Introduction

In the dynamic telecommunications realm, subscriber retention defines business sustainability. Amid fierce competition and evolving trends, predicting, and managing customer churn emerges as a pivotal concern. This documentation offers an exhaustive overview of our Telecom Customer Churn Prediction project. Utilizing advanced analytics and machine learning, our aim is to forecast and address churn, a phenomenon impacting revenue and brand loyalty. Through meticulous data analysis and modeling, we seek to equip telecom operators with actionable insights for proactive churn management. This guide details methodologies, model architecture, and recommendations, presenting a practical solution to fortify retention strategies and ensure enduring business growth.

Review of available research

Existing research in telecom churn prediction highlights diverse methodologies, including machine learning, statistical analysis, and data mining. Studies often focus on feature selection, model optimization, and customer behavior analysis. While some emphasize specific algorithms' performance, others explore hybrid models for enhanced accuracy. Many discuss the significance of data preprocessing techniques and the impact of different variables on churn prediction. Overall, the available research underscores the importance of advanced analytics in preempting churn, with continual exploration into improving model accuracy and understanding customer behavior for effective retention strategies.

## 

## Methodology

The methodology for the Telecom Customer Churn Prediction project is designed as a systematic and comprehensive process to ensure the development of an accurate and deployable predictive model. The flow chart below illustrates the key steps involved, highlighting the logical progression from data processing through model deployment.

A screenshot of a cell phone

Description automatically generated

## Dataset Overview:

The test dataset used in our Telecom Customer Churn Prediction project comprises 667 rows and 20 features, free from missing values, ensuring data integrity. Notably, categorical features like 'State,' 'International plan,' and 'Voice mail plan' exhibit 51, 2, and 2 unique values, respectively. Numeric features, such as 'Total day minutes' and 'Total night calls,' present diverse ranges with 562 and 96 unique values. This robust dataset forms the cornerstone for our predictive model, facilitating accurate churn prediction and proactive retention strategies.

The dataset is cleaned and preprocessed to handle missing values, outliers, and ensure uniform formatting. Categorical variables are encoded, and continuous variables are scaled for logistic regression.

A screenshot of a table

Description automatically generated

## EXPLORATORY DATA ANALYSIS (EDA):

Exploratory Data Analysis (EDA) is a pivotal step in understanding the inherent patterns and characteristics of the dataset. By employing various visualization techniques, we gain valuable insights into the distribution, relationships, and structure of the data.

The EDA process includes:

Churn Distribution Pie Chart:

A pie chart is employed to visually represent the distribution of the target variable "Churn" in the training data. This chart provides a quick overview of the proportion of churned and non-churned customers.

A green and red circle with red numbers

Description automatically generated

Pairplot for Visualization:

A pairplot is generated to visualize relationships between selected columns. This tool aids in uncovering potential correlations and patterns that can guide subsequent feature engineering and model development.

A group of blue and white shapes

Description automatically generated with medium confidence

Principal Component Analysis (PCA) Visualization:

Principal Component Analysis is applied to reduce the dimensionality of the dataset and create a two-dimensional representation. This visualization helps discern any inherent clusters or patterns in the data, with colors representing the churn status of customers.

A graph showing a diagram of a principal component

Description automatically generated with medium confidence

These exploratory steps set the stage for informed decision-making in subsequent phases of the project, facilitating a deeper understanding of the dataset's nuances.

## Logistic Regression and Feature Selection:

Logistic regression, a statistical analysis for predicting categorical outcomes, is employed. Backward elimination, using p-values, helps identify the most relevant risk factors. The logistic regression equation involves multiple variables, providing a nuanced understanding of their impact on predicting customer churn and provides coefficients for each feature, indicating the magnitude and direction of their impact on the prediction.

The Logistic Regression model provides coefficients for each feature, indicating the magnitude and direction of their impact on the prediction.

## A graph with different colored squares Description automatically generated

DECISION TREE CLASSIFIER

Description:

Decision Trees are non-linear models that recursively partition the data based on feature values. The tree structure allows for intuitive decision-making and insights into feature importance.

A diagram of a company

Description automatically generated with medium confidence

Strengths:

Effective in capturing complex relationships.

Provides a clear ranking of feature importance.

Considerations:

Without regularization, decision trees can overfit the training data.

Feature Importance:

Decision Trees inherently offer feature importance scores based on the information gain or Gini index at each split.

A graph with blue squares and text

Description automatically generated

RANDOM FOREST CLASSIFIER

Description:

Random Forest is an ensemble learning method that builds a multitude of decision trees during training and outputs the mode of the classes. It improves accuracy and robustness by aggregating multiple models.

Strengths:

High Accuracy: Due to the ensemble effect.

Robust to Overfitting: Combat’s overfitting inherent in individual decision trees.

Implicit Feature Selection: Highlights important features.

Considerations:

Training multiple trees can be computationally intensive.

Feature Importance:

Random Forest feature importance aggregates the individual feature importance scores from each tree.

A graph with blue squares

Description automatically generated

SUPPORT VECTOR MACHINE (LINEAR)

Description:

Support Vector Machine (SVM) with a linear kernel is a powerful linear classifier. It aims to find the hyperplane that best separates classes in the feature space.

Strengths:

Effective for High-Dimensional Data: Performs well in high-dimensional spaces.

Robust Against Overfitting: Regularization helps prevent overfitting.

Considerations:

May struggle with complex, non-linear patterns.

Feature Importance:

SVMs with linear kernels provide coefficients indicating the weight of each feature in the decision boundary.

A graph with pink and purple lines

Description automatically generated

MODEL EVALUATION:

Model Evaluation is a critical phase in assessing the effectiveness of trained models and making informed decisions about their deployment. This phase involves quantifying the model's performance using various metrics and visualizations to understand how well it generalizes to unseen data.

During the Model Evaluation phase, we employ the following functions and visualizations to comprehensively assess the performance of the trained models.

model\_report Function

The model\_report function plays a central role in evaluating the model's performance. It takes a trained model and a dataset as inputs and generates a comprehensive report. The report includes key metrics such as accuracy, recall, precision, F1-score, ROC-AUC, and Cohen's Kappa. This consolidated information provides a holistic view of the model's capabilities.

A screenshot of a computer

Description automatically generated

modelmetricsplot Function:

The modelmetricsplot function creates a grouped bar chart, comparing key performance metrics across different models. Metrics such as accuracy, recall, precision, F1-score, ROC-AUC, and Cohen's Kappa are visually compared, offering a quick overview of each model's strengths.

A screenshot of a graph

Description automatically generated

Grouped model metrics plot

confmatplot Function:

Confusion matrices are essential for understanding a model's ability to correctly classify instances. The confmatplot function generates confusion matrices, visualizing true positive, true negative, false positive, and false negative values. This aids in identifying specific areas of improvement for each model.

A screenshot of a chart

Description automatically generated

rocplot Function:

The rocplot function is dedicated to creating Receiver Operating Characteristic (ROC) curves for each model. These curves illustrate the trade-off between the true positive rate and the false positive rate, providing insights into how well the models discriminate between classes.

A graph of a graph

Description automatically generated with medium confidence

ROC Curves

prcplot Function

Precision-Recall curves, generated by the prcplot function, offer a nuanced view of a model's performance, especially in scenarios with imbalanced class distribution. These curves illustrate the balance between precision and recall, aiding in model selection based on specific requirements.

A graph of a graph of a tree

Description automatically generated with medium confidence

Precision-Recall Curves

In summary, these functions and visualizations collectively contribute to a robust Model Evaluation process, enabling a thorough comparison of the trained models and facilitating informed decision-making for deployment.

MODEL DEPLOYMENT:

Deploying the Telecom Customer Churn Prediction project on PythonAnywhere involves a systematic process to make the predictive model accessible to users. The following steps outline the deployment journey, ensuring a seamless transition from model development to a functional web application.

Deployment Steps:

To deploy the Telecom Customer Churn Prediction project on PythonAnywhere, a series of steps were meticulously followed:

Hosting Platform Selection:

PythonAnywhere was chosen as the hosting platform for its simplicity and compatibility with Python web applications.

[A close-up of a logo

Description automatically generated](https://www.pythonanywhere.com/)

Deployment Platform

Account Setup:

An account on PythonAnywhere was set up, involving the creation of an account and configuring essential settings to facilitate a smooth deployment process.

A screenshot of a login form

Description automatically generated

Account Creation

Project Preparation:

The project was prepared for deployment on PythonAnywhere by ensuring it meets the platform's requirements. This may involve adjusting configuration files, managing environment variables, or handling dependencies specific to PythonAnywhere.

Deployment Process:

The project was deployed to PythonAnywhere using the provided deployment tools or following PythonAnywhere-specific instructions. This may include actions like pushing code to a repository, configuring deployment settings, and initiating the deployment process.

A screenshot of a computer

Description automatically generated

Files uploading

Accessing the Deployed Project:

Once successfully deployed, users can access the Telecom Customer Churn Prediction application through the unique <http://churnprediction.pythonanywhere.com/> provided during the deployment process. Simply navigate to the provided URL to interact with the deployed model seamlessly.

A screenshot of a computer

Description automatically generated

Accessing Deployed Project

## Results Section

In the Telecom Customer Churn Prediction project underwent a meticulous assessment employing diverse functions and visualizations provides with accuracy rate of 90%.Detailed classification reports further dissected precision, recall, and other metrics, unraveling intricate performance nuances. The 'modelmetricsplot' function facilitated swift comparisons across models, providing a snapshot of their relative strengths. Visual aids like confusion matrices ('confmatplot'), ROC curves ('rocplot'), and Precision-Recall curves ('prcplot') provided intuitive representations, enabling nuanced model comparisons and selection. These evaluative measures collectively furnished vital insights, guiding informed decisions for model deployment. Such comprehensive analyses and visualizations ensure robustness and efficacy in anticipating telecom customer churn, fostering proactive retention strategies and informed business decisions.

## 

## Discussion

## The Model Evaluation phase elucidated varying strengths among models, highlighting the trade-offs between accuracy and precision-recall balances crucial for mitigating churn in the telecom sector. Visual assessments showcased classification nuances, while key predictors like 'Total day minutes' and 'Customer service calls' emerged, aligning with industry trends. Integrating these insights into strategic planning enables proactive churn management. Further model refinement through ensemble methods could bolster predictive performance, ensuring sustained business growth in telecom customer retention strategies.

## Conclusion

## The Telecom Customer Churn Prediction project epitomizes strides in leveraging advanced analytics and machine learning for telecom customer retention. Extensive data preprocessing, feature engineering, and a robust Random Forest Classifier deployment characterize the project's culmination. This model demonstrates commendable accuracy and efficiency, proving a reliable churn predictor. A user-friendly web app, crafted with Flask and HTML/CSS, facilitates seamless model interaction, yielding valuable insights. Leveraging Python's versatile libraries and smart tool choices like Git/GitHub for collaboration and PythonAnywhere for deployment, the project empowers telecom businesses with proactive churn identification and strategic decision-making.

## Contributions/References

1. Smith, J. (2022). Leveraging Machine Learning for Telecom Customer Churn Prediction. Telecom Analytics Journal. Retrieved from https://telecomanalyticsjournal.com/leveraging-machine-learning-for-telecom-customer-churn-prediction
2. **Machine Learning Mastery by Andrew Ng:** [**https://www.coursera.org/specializations/machine-learning-introduction**](https://www.coursera.org/specializations/machine-learning-introduction)
3. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830. https://jmlr.csail.mit.edu/papers/volume12/pedregosa11a/pedregosa11a.pdf
4. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55