Data Analysis using Python

Individual project report submitted in partial fulfillment for the degree of

MSc in Data Analytics

Author:
Bharath Shakthivel - 20057027

April 13th 2025

Table of Contents

1. Introduction

This report explores data analysis using the Python programming language. It provides a brief overview of fundamental data analysis processes such as data cleaning, transformation, and modeling. The primary focus, however, is on performing exploratory data analysis (EDA) on an existing dataset to uncover meaningful insights. Visual analysis is also conducted using various Python libraries and functions. The dataset used for this study is the "Telco-Customer-Churn" dataset, which is analyzed to extract information in both numerical and graphical formats.

1.1 Python for Data Analysis

Python has gained widespread popularity among programmers, especially since its introduction in 1991. It is now considered one of the leading interpreted programming languages, alongside others like Perl and Ruby. Its rise in popularity, particularly after 2005, can be attributed to its utility in building web applications using frameworks such as Django (Python) and Rails (Ruby).

Often referred to as a *scripting language* due to its ease in writing short programs or automating tasks, Python's capabilities extend far beyond that label. While some may view "scripting" as a limitation, Python has proven to be a powerful tool for building serious and scalable software.One of the major reasons for Python's success is its strong support for scientific computing and data analysis. Over the past two decades, Python has evolved from a niche scientific tool to a mainstream language used widely for data science, machine learning, and general-purpose software development across academia and industry.

In the realm of data analysis and visualization, Python is frequently compared with other programming tools and languages such as R, MATLAB, SAS, Stata, and others. Its popularity has been further boosted by the development of powerful open-source libraries like **pandas** and **scikit-learn**, which have made data tasks more accessible and efficient. When combined with its robust capabilities in general-purpose programming, Python stands out as a top choice for building data-driven applications.

1.2 Main Phases in Data Analysis

1. Data Requirements

Data serves as the foundation for any analytical study. The data provided must align with the specific requirements of the analysis. The term *experimental unit* refers to the entity from which data is collected, such as an individual or a group. Key variables from a population—like age,

height, weight, or salary—can be selected regardless of whether they are numerical or categorical in nature.

2. Data Collection

Data collection involves gathering information from various sources based on the goals of the study. This can include relational databases, cloud storage systems, and other digital sources. In some cases, physical sources like field sensors, traffic cameras, satellites, and monitoring devices are also used to acquire data.

3. Data Processing

Once collected, data needs to be organized for analysis. This step typically includes formatting the data into structured forms, such as tables with rows and columns. Spreadsheet applications and statistical tools are often used during this stage to prepare the data for further exploration.

4. Data Cleaning

After organizing the data, the next step is cleaning it to eliminate inconsistencies, errors, and duplicates. This process ensures data accuracy and quality. Common cleaning tasks include matching records, identifying inaccuracies, sorting data, detecting outliers, correcting spelling in textual data, and maintaining overall data integrity. Clean data is crucial to obtaining reliable and meaningful results.

5. Exploratory Data Analysis (EDA)

Once the dataset is cleaned, exploratory data analysis can be conducted. This phase involves uncovering patterns and insights using techniques like descriptive statistics (e.g., mean, median) and visualizations (e.g., graphs, charts). Visualization tools help reveal trends and relationships within the data, offering a clearer understanding of the information.

6. Modeling and Algorithms

Mathematical models or algorithms are then applied to the data to explore relationships between variables. These models can help identify correlations, causal effects, or other meaningful patterns, forming the basis for predictive or inferential analysis.

7. Data Product

The final phase often involves creating a *data product*—a software or system that takes processed data as input and produces useful outputs. These outputs may be based on algorithms or models and can be integrated into larger systems or applications to support decision-making or automation.

1.3 Dataset Selection

With the rapid growth of the telecommunications industry, service providers are increasingly focused on expanding their subscriber base. However, in today's highly competitive market, retaining existing customers has become a significant challenge. Studies in the telecom sector have shown that acquiring new customers is considerably more expensive than retaining current ones. As a result, leveraging insights from customer data can help predict whether a customer is likely to stay or leave the company. Based on these predictions, telecom companies can take proactive measures to retain their customers, thereby maintaining their market position and long-term value.

1.4 Libraries and Tools Used for Data Analysis

1. Pandas

Pandas is a powerful open-source library designed for data manipulation and analysis. It provides two main data structures: Series (1D) and DataFrame (2D), which allow users to easily load, filter, clean, and transform structured data. It's particularly useful for tasks like handling missing data, merging datasets, and performing group-wise operations.

2. NumPy

NumPy (Numerical Python) is the core library for numerical computing in Python. It offers support for multi-dimensional arrays and a wide range of mathematical operations, making it essential for handling large numerical datasets. Many other libraries, including pandas and scikit-learn, rely on NumPy under the hood.

3. Matplotlib

Matplotlib is a fundamental data visualization library in Python. It enables users to create static, animated, and interactive plots, such as line graphs, bar charts, histograms, and scatter plots. It provides a high degree of control over every aspect of a figure, making it ideal for detailed and customizable plots.

4. Seaborn

Seaborn is a high-level visualization library built on top of matplotlib. It simplifies the creation of statistical and attractive visualizations, including heatmaps, box plots, and violin plots. Seaborn works seamlessly with pandas DataFrames and offers improved default styling and color schemes.

5. Jupyter Notebook

Jupyter Notebook is an open-source web-based tool that allows users to create and share documents that contain live code, equations, visualizations, and narrative text. It is widely used in data science and academic research for documenting analysis workflows. It supports

interactive data exploration, visual outputs, and step-by-step code execution, making it an ideal environment for data analysis and experimentation.

2. Literature Review

2.1 Real-World Example (Customer Churn Prediction in Telecom Industry)

The article *"Customer Churn Prediction in Telecommunication Industry: A Data Analysis Techniques Approach"* delves into the significance of predicting customer churn within the telecom sector. It emphasizes the application of data analysis techniques to identify patterns and factors contributing to customer attrition.

Key Challenges

- High Competition: The telecom industry faces intense competition, making customer retention a critical concern.

- Data Complexity: Handling vast amounts of customer data, including call records, service usage, and billing information, poses challenges in data processing and analysis.

Data Analysis Techniques

- Data Preprocessing: Involves cleaning and organizing data to ensure accuracy and consistency.

- Feature Selection: Identifying relevant variables that significantly impact customer churn.

- Predictive Modeling: Utilizing machine learning algorithms to forecast potential churners based on historical data.

Summary

The study underscores the importance of leveraging data analysis techniques to proactively identify customers at risk of churning. By implementing predictive models, telecom companies can develop targeted strategies to enhance customer retention and maintain a competitive edge in the market.

2.2 Wes McKinney – *Python for Data Analysis (O'Reilly Media)*

A key reference in the field of data analysis using Python is *Python for Data Analysis* by Wes McKinney, published by O'Reilly. This book serves as a comprehensive guide to understanding data manipulation, preparation, and exploration using the Python programming language.

Wes McKinney, the creator of the pandas library, provides readers with a practical and detailed approach to working with real-world datasets. The book covers essential tools and techniques, including data wrangling, aggregation, and visualization using libraries like pandas, NumPy, matplotlib, and more.

What makes this book particularly valuable is its beginner-friendly yet thorough explanation of core data analysis workflows. It provides step-by-step coding examples that help users not only understand the syntax but also the logic behind analytical decisions. The chapters on handling time series data and working with missing or messy data were especially helpful during our project's preprocessing phase.

This resource greatly enhanced our ability to perform efficient exploratory data analysis and build a clean foundation before applying any machine learning models.

2.3 Sharmila K. Wagh – Customer Churn Prediction in Telecom Sector Using Machine Learning Techniques

This study explores the development of a machine learning-based system for predicting customer churn in the telecom industry. The approach is centered around building a robust and interpretable model that can identify potential churners early, allowing telecom providers to take proactive retention measures.

Key Focus Areas

- **Machine Learning Integration**: The system leverages classification algorithms, particularly Decision Tree and Random Forest, to classify customers as churners or non-churners based on behavioral and service usage patterns.

- **Data Preprocessing**: The raw data undergoes thorough cleaning and preparation, including handling missing values, removing redundant features, and converting data types to ensure optimal model input.

- **Feature Selection**: Relevant features are selected using correlation-based methods to improve model efficiency and accuracy. This step ensures that only the most informative

attributes are retained for training.

- **Handling Imbalanced Data**: Techniques such as SMOTE (Synthetic Minority Over-sampling Technique) and ENN (Edited Nearest Neighbors) are applied to address the challenge of class imbalance, which is common in churn datasets.

- **Survival Analysis Techniques**: In addition to classification models, the study incorporates survival analysis and the Cox Proportional Hazard model to estimate the likelihood and timing of churn. This allows the system to not only predict if a customer will churn but also when it is most likely to occur.

- **Visualization and Interpretation**: The approach includes both univariate and bivariate analysis to explore relationships between features and churn behavior. Important features are ranked to provide interpretability and guide business decisions.

Summary

The paper presents a hybrid approach combining classical machine learning models with statistical survival analysis techniques. It emphasizes the need for domain-specific preprocessing, effective feature selection, and targeted modeling strategies. By integrating classification algorithms with time-to-event analysis, the system offers a well-rounded solution for understanding and mitigating customer churn in the telecom sector.

3. Methodology

This study follows a systematic approach to understanding and predicting customer churn in the telecommunications industry. The methodology comprises multiple phases, from data acquisition to statistical modeling and evaluation.

3.1 Data Collection and Preprocessing

The dataset used for this study consists of 7,043 customer records from a telecommunications company, each characterized by 21 features including demographic information, account details, service subscriptions, and churn status. The dataset was imported using **Pandas**, and initial data inspection revealed the presence of categorical and numerical variables, along with a few inconsistencies such as missing or incorrect data types.

Preprocessing steps included:

- Dropping the non-informative customerID field.
- Converting TotalCharges from object to float using pd.to_numeric(), which revealed 11 missing values. These rows were removed, resulting in a cleaned dataset of 7,032 entries.
- Ensuring there were no duplicate records.
- Transforming categorical variables into numerical format using one-hot encoding to prepare the data for modeling.
- Addressing missing values through imputation techniques where applicable, although in this case, rows with missing TotalCharges were dropped for simplicity.

3.2 Exploratory Data Analysis (EDA)

EDA was conducted using **Seaborn** and **Matplotlib** to understand patterns and relationships within the data. Key techniques included:

- **Univariate analysis** through histograms and pie charts to understand feature distributions (e.g., churn proportion, tenure, charges).
- **Bivariate analysis** using boxplots and count plots to evaluate relationships between categorical/numerical features and the target variable.
- **Multivariate analysis** using pair plots and grouped bar charts to explore interactions among multiple features.
- **Correlation matrix** to identify highly correlated features such as TotalCharges and tenure.

This phase uncovered several important patterns, including high churn among customers with short tenures, those on month-to-month contracts, and those using electronic check as a payment method.

3.3 Feature Engineering

Categorical features, such as InternetService, Contract, and TechSupport, were transformed via one-hot encoding. Binary features (e.g., Yes/No, Male/Female) were converted to numerical values for modeling purposes.

3.4 Predictive Modeling

A **Logistic Regression** model was employed to identify the key drivers of churn. The model was chosen for its interpretability and ability to estimate the probability of customer churn. Feature scaling was not required due to the nature of the model and data transformation.

Model training was conducted using the processed features as predictors (X) and the churn outcome (y) as the target variable. The model was then evaluated using confusion matrices to assess its classification performance.

4. Results

4.1 Descriptive Findings

The analysis revealed several critical insights regarding customer churn behavior:

- Approximately **26.5%** of customers had churned.
- **Short tenure** (especially within the first 10 months) was a strong indicator of churn.
- Customers with **monthly charges above $70** exhibited a higher churn rate.
- The majority of churned customers had:
    - Month-to-month contracts
    - Paperless billing
    - Electronic check payment methods
    - No tech support, online security, or device protection

4.2 Visual Insights

- **Histograms and pie charts** highlighted tenure and monthly charges as influential numerical variables.
- **Count plots** emphasized that customers without dependents, partners, or service add-ons were more prone to churn.
- **Line plots** showed a relationship between high monthly charges and shorter tenure with higher churn probability.
- **Heatmaps** illustrated correlations among features, notably between TotalCharges, MonthlyCharges, and tenure.

4.3 Model Outcomes

The logistic regression model provided interpretable coefficients for each feature:

- Positive coefficients (increased likelihood of churn) were found for features such as **Electronic Check**, **Month-to-Month Contract**, and **Fiber Optic Internet**.
- Negative coefficients (reduced churn likelihood) were associated with **Tech Support**, **Online Backup**, and **Two-Year Contracts**.

A **confusion matrix** was used to evaluate model performance, showing reasonably balanced predictive power. A normalized confusion matrix revealed that the model was more accurate in identifying non-churners than churners, suggesting class imbalance—a common issue in churn prediction problems.

5. Conclusion and Recommendations

5.1 Conclusion

This study investigated the determinants of customer churn within the telecommunications sector through a comprehensive data-driven approach. By employing exploratory data analysis, categorical variable encoding, and logistic regression modeling, the project successfully identified patterns and predictors associated with customer attrition.

The findings indicate that churn is significantly influenced by contract type, payment method, support services, and monthly charges. Specifically, customers subscribed to **month-to-month contracts**, those using **electronic check** as their payment method, and those not availing themselves of **technical support** or **online services** demonstrated a markedly higher likelihood of discontinuing their services. Furthermore, a notable proportion of churn events occurred within the **first 10 months** of tenure, suggesting that customer satisfaction and engagement in the early stages of service are critical to retention.

The application of logistic regression provided interpretable coefficients, facilitating an understanding of the relative importance of each predictor. Although the model displayed stronger performance in predicting non-churn instances than churn, it offered valuable insights into the behavioral and demographic characteristics of at-risk customers.

5.2 Contributions and Learning Outcomes

This project contributed to both practical and theoretical domains. Practically, it delivered actionable insights that telecommunications firms can leverage to reduce churn and enhance customer retention strategies. Theoretically, it reinforced the applicability of logistic regression in binary classification problems and demonstrated the effectiveness of visual and statistical exploration in uncovering patterns in high-dimensional data.

The project also contributed to the author's academic and professional development by enhancing skills in:

- Data preprocessing and cleaning
- Exploratory data analysis and visualization
- Feature transformation and encoding
- Predictive modeling using logistic regression
- Interpretation of model outcomes for business decision-making

5.3 Recommendations

Based on the insights obtained, the following recommendations are proposed to mitigate churn:

1. **Strengthen Early Engagement**: Given the high churn rate among new customers, firms should implement structured onboarding programs and personalized engagement strategies within the first few months of service.

2. **Encourage Long-Term Contracts**: Incentivizing customers to opt for one- or two-year contracts could lead to reduced churn rates compared to flexible month-to-month arrangements.

3. **Enhance Support Services**: Investment in technical support, online security, and data backup services can add value and increase customer satisfaction, particularly among fiber-optic users who exhibit higher churn rates.

4. **Review Billing and Pricing Structures**: Since higher monthly charges correlate with increased churn, companies should evaluate pricing fairness and provide flexible billing options or bundled packages.

5.4 Future Research Directions

Several limitations of this study present opportunities for future work. These include:

- **Model Expansion**: Future research should explore more advanced classification algorithms such as Random Forests, Gradient Boosting Machines (GBM), or ensemble methods to improve prediction performance.

- **Addressing Class Imbalance**: Techniques such as SMOTE (Synthetic Minority Over-sampling Technique), resampling, or penalized models should be incorporated to better handle the inherent class imbalance.

- **Temporal and Sequential Analysis**: Incorporating time-based variables and conducting survival analysis could offer a deeper understanding of customer lifecycles and churn timing.

- **Customer Segmentation**: Cluster analysis or unsupervised learning could be used to develop customer personas and support the design of targeted retention campaigns.

- **Deployment of Real-Time Systems**: Implementing real-time analytics platforms would allow for proactive churn prevention through timely alerts and personalized interventions.

6. References

1. Cummings, T. (2021). *Hands-On Data Analysis with Pandas: Efficiently Perform Data Collection, Wrangling, Analysis, and Visualization Using Python*. Packt Publishing.
   Available at Google Books

2. Yadav, A. (2018). *Data Analysis with Python: A Modern Approach*. BPB Publications.
   Available at Google Books

3. Balaji, C., & Uma, G. (2021). *Data Analysis Using Python*. *International Journal of Engineering Research & Technology (IJERT)*, 10(7), 240–246.
   Available at ResearchGate/Cloudfront

4. Albon, C. (2015). *Data Analysis and Visualization with Python*. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 1912–1917). IEEE. https://doi.org/10.1109/BigData.2015.7359318
   Available via IEEE Xplore

5. Khan, M. M., Taneja, M., & Ray, P. P. (2023). Data analytics using Python for smart healthcare monitoring. *Software Impacts*, 15, 100494. https://doi.org/10.1016/j.simpa.2023.100494
   Available at ScienceDirect