**Customer Churn Prediction System with Data Integration for Telcom Company**

**Martin Gitau Kimani**

**100994**

**An Information System Project 2 submitted to the School of**

**Computing and Engineering Sciences in partial fulfilment of the requirements**

**For the award of the bachelor’s degree in Business Information Technology of**

**Strathmore University**

**School of Computing and Engineering Sciences**

**Strathmore University**

**Nairobi, Kenya.**

**April 2024**

# Declaration and approval

I hereby declare that this work has not been previously submitted and approved by this or any

Other University. To the best of my knowledge and belief, the report contains no material

Previously published or written by another person except where due reference is made in the

Research project itself.

Student’s signature

Sign: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Supervisor’s signature

Sign: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Abstract

In the contemporary information technology landscape, data has evolved into a valuable asset, often likened to the significance of oil. Recognizing data's pivotal role in organizational success, effective utilization has become imperative for staying competitive and creating value. However, the raw nature of data requires transformation into meaningful insights, a task typically undertaken by professionals such as Data Analysts, BI Analysts, and Data Scientists. This underscores the essential need for robust data integration within businesses to extract actionable insights.

Customer churn, the phenomenon where consumers disengage from online products or services, poses a significant challenge for businesses, resulting in reduced revenue and sluggish growth. While this phenomenon may lack a straightforward explanation, leveraging data can provide crucial insights into understanding and predicting customer churn. This emphasizes the central focus of ensuring proper data integration in business operations, utilizing data to not only comprehend the issue but also to anticipate and mitigate the occurrence of customer churn. Creating a data eco system that allows for the support a customer churn predictive system is paramount.

Table of content

[Declaration and approval ii](#_Toc163821398)

[Abstract iii](#_Toc163821399)

[Table of content iv](#_Toc163821400)

[List of Figures viii](#_Toc163821401)

[Chapter 1: Introduction 1](#_Toc163821402)

[1.1 Background 1](#_Toc163821403)

[1.2 Problem Statement 2](#_Toc163821404)

[1.3 Objectives 3](#_Toc163821405)

[1.3.1 General Objective 3](#_Toc163821406)

[1.3.2 Specific objectives 3](#_Toc163821407)

[1.4 Research Questions 3](#_Toc163821408)

[1.5 Justification 3](#_Toc163821409)

[1.6 Scope and limitations 4](#_Toc163821410)

[Chapter 2: Literature Review 5](#_Toc163821411)

[2.1 Introduction 5](#_Toc163821412)

[2.2 Problem of Data integration 5](#_Toc163821413)

[2.2.1 State of data integration 5](#_Toc163821414)

[2.2.2 Data integration methods and techniques 7](#_Toc163821415)

[2.2.3 Data Warehousing 7](#_Toc163821416)

[2.2.4 Causes of customer churn 8](#_Toc163821417)

[2.2.5 Customer purchase behavior 8](#_Toc163821418)

[2.2.6 Predictive analysis 9](#_Toc163821419)

[2.2.7 Develop Customer Churn Predictive Analysis and implementing data pipelines 10](#_Toc163821420)

[2.3 Related works 12](#_Toc163821421)

[2.3.1 Machine Learning for Environmental Data Integration in Conservation Initiatives 12](#_Toc163821422)

[2.3.2 DataOps 13](#_Toc163821423)

[2.3.3 Customer churn prediction in telecom using machine learning in big data platform 13](#_Toc163821424)

[2.4 Gaps Customer churn with data integration 15](#_Toc163821425)

[2.5 Conceptual Frame work 16](#_Toc163821426)

[2.5.1 Generation 16](#_Toc163821427)

[2.5.2 Ingestion 17](#_Toc163821428)

[2.5.3 Transformation 17](#_Toc163821429)

[2.5.4 Storage 18](#_Toc163821430)

[2.5.5 Serving for Machine Learning / Data Analytics: 19](#_Toc163821431)

[2.5.6 Underlying Factors 20](#_Toc163821432)

[Chapter 3: Methodology 22](#_Toc163821433)

[3.1 Introduction 22](#_Toc163821434)

[3.2 Development Approach 22](#_Toc163821435)

[3.3 Justification of the Methodology 22](#_Toc163821436)

[3.4 System Analysis 22](#_Toc163821437)

[3.4.1 Context Diagram 23](#_Toc163821438)

[3.4.2 DFD Level 0, 1 23](#_Toc163821439)

[3.4.3 ERD (Entity-Relationship Diagram) 24](#_Toc163821440)

[3.4.4 Database Schema 24](#_Toc163821441)

[3.5 System Design 24](#_Toc163821442)

[3.6 System Deliverable/Milestones 25](#_Toc163821443)

[Deliverables 25](#_Toc163821444)

[Milestones 26](#_Toc163821445)

[3.6.1 System proposal 27](#_Toc163821446)

[Chapter 4: System Analysis and Design 28](#_Toc163821447)

[4.1 Introduction 28](#_Toc163821448)

[4.2 System Requirements/Requirements Gathering 28](#_Toc163821449)

[4.2.1 Functional Requirements 29](#_Toc163821450)

[4.2.2 Non-Functional Requirements 29](#_Toc163821451)

[4.3 System Analysis Diagrams 30](#_Toc163821452)

[4.3.1 Context Diagrams 30](#_Toc163821453)

[4.3.2 Data Flow Diagrams (Level 0, Level 1) 30](#_Toc163821454)

[4.3.3 Activity Diagrams 32](#_Toc163821455)

[4.4 System Design Diagrams 34](#_Toc163821456)

[4.4.1 Entity-Relationship Diagrams (ERD) 34](#_Toc163821457)

[4.4.2 Logical Database Schema 35](#_Toc163821458)

[Chapter 5: System Implementation and testing 36](#_Toc163821459)

[5.1 Introduction 36](#_Toc163821460)

[5.2 Description of the Implementation Environment 36](#_Toc163821461)

[5.2.1 Hardware Specifications 36](#_Toc163821462)

[5.2.2 Software Specifications 37](#_Toc163821463)

[5.3 Description of the Dataset 37](#_Toc163821464)

[5.4 Airflow and work orchestration 39](#_Toc163821465)

[5.5 Description of Testing 40](#_Toc163821466)

[5.5.1 Exploratory data analysis 40](#_Toc163821467)

[5.5.2 Preparation for machine model building 44](#_Toc163821468)

[5.5.3 Model building 46](#_Toc163821469)

[5.5.4 Model saving and storage 49](#_Toc163821470)

[5.6 Testing results and verdict 50](#_Toc163821471)

[Chapter 6: CONCLUSIONS, RECOMMENTATIONS AND FUTURE WORKS 54](#_Toc163821472)

[6.1 Conclusions 54](#_Toc163821473)

[6.2 Recommendations 54](#_Toc163821474)

[6.3 Future works 54](#_Toc163821475)

[References 55](#_Toc163821476)

[Appendix 57](#_Toc163821477)

List of Figures

[Figure 1.1 Data Engineering life cycle 2](#_Toc164437325)

[Figure 2.1 Components of ETL process 7](#_Toc164437326)

[Figure 2.2 Conceptual Frame work 16](#_Toc164437327)

[Figure 4.1 Context 30](#_Toc164437328)

[Figure 4.2 DFD level 1 31](#_Toc164437329)

[Figure 4.3 Flow chart 32](#_Toc164437330)

[Figure 4.4 ERD diagram 33](#_Toc164437331)

[Figure 4.5 Database schema 34](#_Toc164437332)

[Figure 5.1 ETL code 37](#_Toc164437333)

[Figure 5.2 Airflow interface 38](#_Toc164437334)

[Figure 5.3 Ratio of churned to non-churned customers 39](#_Toc164437335)

[Figure 5.4 Female and male churning 40](#_Toc164437336)

[Figure 5.5 Senior Citizen churning rate 41](#_Toc164437337)

[Figure 5.6 Contract type in relation to churning 42](#_Toc164437338)

[Figure 5.7 Snippet of data structure 43](#_Toc164437339)

[Figure 5.8 Correlation between features 44](#_Toc164437340)

[Figure 5.9 Decision Tree Visualization 45](#_Toc164437341)

[Figure 0.1 Gant Chart 56](#_Toc164437342)

# Introduction

## Background

Data integration is a big part of businesses as data has become a key factor in business growth. It may seem that every business should jump onto the Data analytics and Data science band wagon in order to tap into insights they produce, but this can be detrimental and may yield insufficient results (Joe Reis, 2022). This is because data integration is important for businesses and depending on the level of business or organization its data maturity stage may differ. Data maturity stages may vary. These stages are; starting with data, Scaling with data and Leading with data from start-ups to big companies respectively (Joe Reis, 2022). So this means it is very important for data engineers to know at what stage of maturity the organization is for proper data integration. For this scenario we looked at an ecommerce problem of customer churn.

Data engineering can be defined as implementation and maintenance of systems that allow for raw data intake and production of consistent information that is of high quality that supports downstream use cases such as analysis and machine learning, also alongside this it would involve data management, DataOps and software engineering (Joe Reis, 2022). All this is essential for the integration of data in any company. Nowadays it is seen that data analyst and data scientist are spending most of their time around 70% and 80% of their time cleaning and organizing data. This is due to failure of implementing proper data infrastructure from immature data engineering and data science practices (Ralph Kimball, 2019).

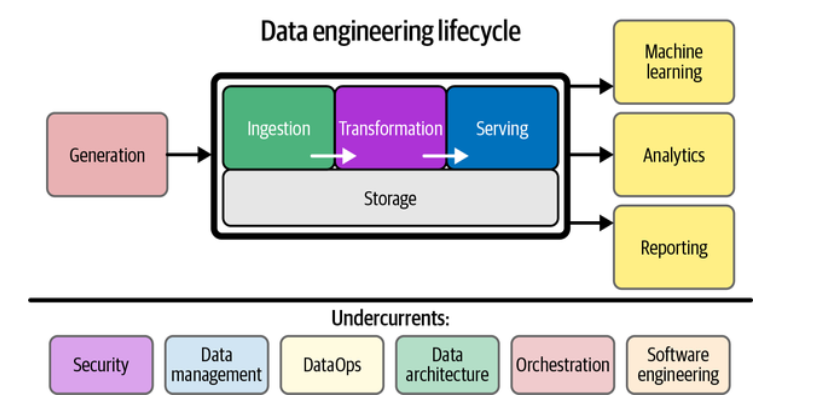
This can be approached by implementing and focusing on the data engineering life cycle that deals with the data engineering lifecycle. Let’s for example a cosmetics company that has thousands of users and are active on social media. The company draw data from the various social media platforms, store them and serves the data for downstream use cases then provide a visualization with tools like power BI to provide a live feed of their product and how it is performing based on what users are saying about it on social media. The project aimed to underscore the pivotal role of data integration within companies, with a specific focus on addressing the customer churn challenge. Our approach involved implementing a robust data pipeline designed to seamlessly gather data from diverse sources, transform it into a structured format, and conduct thorough exploratory and analytical processes. Leveraging machine learning techniques, the project aimed to predict customer churn, highlighting how strategic data integration plays a crucial role in deriving actionable insights for informed decision-making.

Figure 1.1 Data Engineering life cycle

## Problem Statement

Customer churn, the process where customers cease their use of a product or service, is a significant concern for businesses as it directly impacts revenue and growth. Recognizing and addressing customer churn is essential for maintaining a prosperous business. In this context, data integration plays a crucial role. It involves the seamless merging and synchronization of data from various sources, offering a comprehensive perspective of customer interactions, behaviors, and preferences.

This consolidated data serves as the basis for applying advanced analytics and machine learning algorithms. With a solid data integration framework, predictive analytics can be utilized to foresee customer churn. This is achieved by examining past customer data and pinpointing patterns or signs that lead to churn, enabling organizations to take proactive steps to retain valuable customers. Effective data integration facilitates real-time tracking and analysis of customer interactions.

By streaming data from a variety of sources, such as user engagement on platforms or social media sentiments, organizations can keep up with changing customer sentiments. This dynamic insight, integrated through data pipelines, enables quick decision-making in response to emerging churn risks. Data integration also tackles the issue of inaccurate insights resulting from disparate data sources. By unifying data through a well-coordinated integration process, organizations can ensure the accuracy and reliability of the insights derived, thereby improving the effectiveness of existing customer churn prediction models. In essence, data integration is a key component in proactively addressing customer churn. By leveraging a unified data landscape, organizations can not only understand the intricacies of customer behavior but also use advanced analytics to predict and preemptively respond to potential churn. This not only fosters customer retention but also promotes business sustainability. The current advancements in data integration and predictive analytics are enhancing and supporting the existing customer churn prediction models and contributing to more effective model building.

## Objectives

### General Objective

Creating data pipelines involves efficiently collecting and preprocessing diverse data sources for the subsequent creation and execution of customer churn prediction models

### Specific objectives

1. To investigate data integration methods, techniques and investigating challenges in its implementation
2. To analyze customer purchase behaviors and data in online services
3. To design and develop a predictive model for customer churn using a functional streaming data pipeline
4. To test the functionality and effectiveness of the customer churn predictive model by evaluating its performance within the streaming data pipeline.

## Research Questions

1. How do various factors contribute to customer churn within the industry
2. In which way do various factors influence customer purchasing decisions
3. How can we address challenges to effective data integration
4. How are various data science techniques and algorithms used in predictive analysis
5. How can we implement current technologies to implement data pipelines

## Justification

In the absence of an automated system that are not integrated within the data eco-system of a company in the context of customer churn and facilitating seamless data integration, business is left vulnerable to significant risks. Extensive research by (Argelaguet, 2021) has shown that companies without robust predictive analytics systems that are not integrated within the organization using data ETL pipelines find it difficult in addressing the issue of applying analytical and machine learning techniques efficiently. Also, the lack of efficient data integrating solution can lead to data silos that hinder critical decision making processes. This greatly undermines the organization’s competitiveness

## Scope and limitations

This only covered the data pipelines for cases of medium sized companies and did not include big data and IOT devices. It’s possible to integrate big data and data from IOT devices in future but for case it was focusing on small to medium ecommerce business.

# Literature Review

## Introduction

In this chapter we look at some of the existing body of knowledge in the field of data integration and data engineering. The aim of the chapter is to provide a more comprehensive understanding of the current state of data engineering and data integration along with the challenges it presents. We also looked at the current solutions that have been put in place or proposed. Customer churn was another problem we looked at and tried to understand it much better as it was our downstream use-case that allowed to see how the whole process of data engineering fits in and how effective it can be in developing predictive system for customer churn.

## Problem of Data integration

The process of data integration, which involves extracting data from various sources and transforming it into a standardized and analyzable format, can be fraught with challenges. In many cases, there are no established best practices for efficient data integration, leading to a need for more effective processes (Hira, 2023). Industries and research institutions often find that their data storage infrastructure is insufficient to handle the volume and variety of data they generate. This necessitates advancements in workflows, including data storage, data management, data maintenance, data integration, and data interoperability (Joe Reis, 2022).

Among these, data integration and data interoperability are of particular importance for organizations seeking to enhance their workflows. However, these challenges can become particularly complex for companies or organizations planning to implement big data architectures. This is because the data they work with is often heterogeneous in nature (J. M., 2021). The complexity of data integration is further exacerbated by the need to integrate data from various sources, each with its own unique characteristics and formats. This can lead to issues with data consistency, quality, and reliability, which can in turn impact the effectiveness of data-driven decision-making processes (Reis, 2022).

Despite these challenges, advancements in data integration technologies and practices continue to evolve, providing organizations with more effective tools and methods for managing and leveraging their data (Reis, 2022).

### State of data integration

Currently, many companies may not fully grasp the vital importance of data integration in their data analytics and data science endeavors. This oversight becomes apparent when organizations hastily adopt popular data analysis and data science practices, believing they yielded immediate benefits. In such cases, a disconcerting pattern emerges where data scientists and analysts within these companies, which have not integrated robust data integration practices, find themselves disproportionately occupied with the arduous task of data cleaning and reformatting before they can even commence meaningful analytical computations (Murphy, 2022).

This prevailing scenario underscores a profound issue: the failure to appreciate data integration's foundational role in the data ecosystem. It is not merely a preparatory step; it is the linchpin upon which the edifice of data-driven insights and informed decision-making rests. Without adequate data integration, data remains a disparate and unwieldy assortment, a far cry from the harmonious, structured dataset necessary for sound analytics. Consequently, organizations inadvertently squander substantial resources, both in terms of time and human capital, on the avoidable "data wrangling" stage.

The consequences of neglecting data integration are manifold. Not only does it impede the pace of data-driven projects, but it also compromises the quality of insights and, by extension, the accuracy of decision-making. The value that data scientists and analysts bring to the table is undercut, as they find themselves bogged down by data-related drudgery rather than focusing on the core analytics and innovation.

Therefore, it is imperative for organizations to recognize that data integration is not an optional component but a fundamental prerequisite for meaningful data utilization. By embracing robust data integration practices, businesses can streamline their data workflows, unlock the full potential of their data assets, and empower their data experts to engage in value-adding activities. In doing so, they position themselves to harness the true benefits of data analytics and data science.

As of the moment most companies are not aware how crucial data integration is. This can be seen in instances where companies jump on the band wagon of the most popular data analysis and data science practices that have been seen to have very beneficial for some organization. What happned is that the data scientist and analysts that are employed in these organization which have not incorporated any kind of data integration practices spend most of their time trying to clean the data and format it to proper structure before performing any analytical calculation on them

### Data integration methods and techniques

Understanding effective data integration techniques is crucial for the development of efficient data pipelines. There are two primary approaches to achieve this

#### ETL

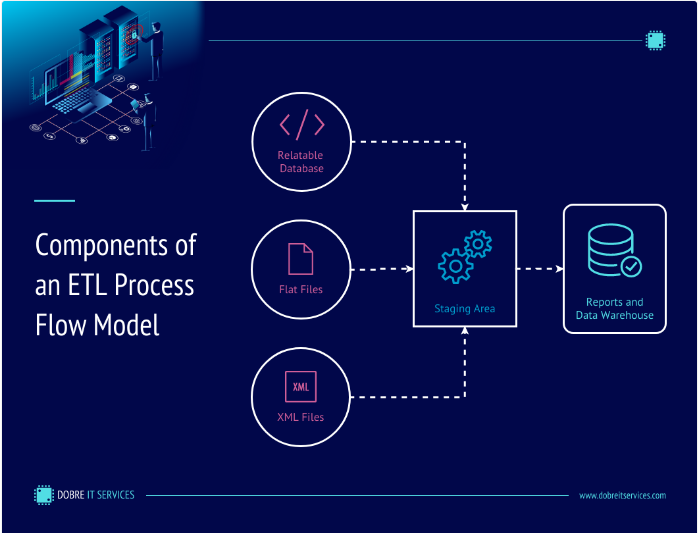


Figure 2.1 Components of ETL process

First we have ETL Processes, which stands for Extract, Transform, Load, are fundamental processes in data integration. This involves the extraction of raw data from source systems, transforming it into a usable format, and loading it into the target system. Tools like Talend and Informatica are widely employed for ETL processes. These tools streamline the data transformation journey, making data ready for analysis and other downstream operations.

### Data Warehousing

Secondly we have Data warehousing where it focuses on centralizing data from various sources into a data warehouse. This centralized repository is designed for analysis and reporting purposes. Solutions like Amazon Redshift and Snowflake are popular choices for data warehousing. These platforms offer scalability and high performance, making them well-suited for handling and analyzing large volumes of data (Bhatia, 2019). These two approaches, ETL processes and data warehousing, are critical components of efficient data integration strategies. They enable organizations to consolidate, transform, and analyze data from diverse sources, empowering data-driven decision-making and insights (Bhatia, 2019)

### Causes of customer churn

Understanding the factors contributing to customer churn is a critical endeavor in enhancing customer retention. Various methodologies have been employed to delve deeper into this issue. One such approach involves conducting customer surveys, where companies and organizations gather valuable feedback from their clientele. These surveys serve as a means to identify pain points and unearth the underlying reasons that drive customers away. By carefully analyzing the responses, recurring themes and issues can be uncovered, offering actionable insights for improvement (Çelik, 2019).

Another method, cohort analysis, plays a pivotal role in comprehending customer behavior. This technique involves segregating customers into distinct groups based on specific characteristics, such as when they signed up or how they use a product (Fedushko, 2022). By examining these cohorts, organizations can identify patterns associated with churn. This data-driven approach provides a means to understand why certain customer groups are more susceptible to churn, enabling proactive measures to be taken to address the issue. These methods collectively empower organizations to gain insights into customer churn, fostering the development of effective strategies for customer retention and overall business enhancement.

### Customer purchase behavior

This endeavor involves a detailed analysis of customer purchasing behavior in online platforms, scrutinizing various elements like user interactions, buying patterns, and conversion rates. There are established methods to tackle this challenge effectively. One approach leverages tools such as Google Analytics and Mix panel, which provide valuable insights into user behavior on websites and mobile applications. These tools enable organizations to gain a comprehensive understanding of how users interact with their platforms, identify trends in purchasing behavior, and assess the efficacy of conversion strategies (McGuirk, 2023).

Another solution in the arsenal is A/B testing, a method to evaluate the impact of changes on user behavior. This is achieved by comparing the performance of two or more versions of webpages or apps. Organizations can systematically analyze which design or content elements lead to improved conversion rates, allowing for data-driven decisions in enhancing user experiences and achieving higher conversion rates (Yu, 2020). These strategies collectively empower businesses to optimize their online platforms and bolster customer engagement.

Within the existing strategies for analyzing customer behavior on online platforms, a critical gap lies in the absence of a dedicated mechanism for predicting and addressing customer churn. The challenge of foreseeing potential customer departures before they occur is a crucial aspect that hasn't been adequately addressed in the current framework. Our project seeks to bridge this gap by introducing a customer churn prediction system integrated with ETL (Extract, Transform, Load) pipelines. These pipelines facilited the seamless collection, transformation, and loading of data, enabling organizations to harness valuable insights. By leveraging historical data and employing predictive analytics, our system aims to proactively identify patterns and indicators that precede customer churn.

The integration of ETL pipelines ensures that the predictive models are fueled with accurate and timely data, enhancing their efficacy. This comprehensive solution empowers organizations to not only anticipate customer churn but also implement targeted strategies to retain valuable customers. In essence, our project endeavors to enhance the existing framework by providing organizations with a proactive and integrated approach to customer churn mitigation.

### Predictive analysis

In this phase, the project harnesses advanced machine learning algorithms, including decision trees, logistic regression, and random forests, to delve deep into the data and make accurate predictions, particularly focused on customer churn. These algorithms excel in identifying intricate patterns and relationships within the data, providing a foundation for highly precise predictions (Fedushko, 2022). For instance, decision trees dissect data into branches of conditions, logistic regression evaluates the probability of a specific event like customer churn, and random forests aggregate the power of multiple decision trees to enhance accuracy. This ensemble of algorithms collectively forms a robust framework for making informed predictions regarding customer churn (Fedushko, 2022).

Additionally, the project employs statistical modeling techniques such as survival analysis and regression models to gain comprehensive insights into the factors influencing the likelihood of customer churn (Fedushko, 2022). Survival analysis, specifically adept at assessing the time to an event, is instrumental in understanding when and why customers are likely to churn. On the other hand, regression models establish relationships between various factors and the propensity for customer churn, further enhancing the project's ability to provide insightful predictions (Fedushko, 2022).

In the context of the project's focus on customer churn prediction, these sophisticated algorithms and statistical modeling techniques contribute to a nuanced understanding of the variables influencing customer behavior. By leveraging these tools, the project aims to provide organizations with actionable insights that go beyond simple predictions, enabling them to implement targeted strategies for customer retention effectively.

### Develop Customer Churn Predictive Analysis and implementing data pipelines

In the endeavor to implement comprehensive data pipelines and cater to downstream use cases, organizations often depend on data integration platforms and tools as linchpins of their data strategy. These platforms play a multifaceted role in the data lifecycle and data engineering lifecycle, particularly in streamlining the Extract, Transform, Load (ETL) processes, and ensuring that data is not only efficiently processed but also readily accessible for analysis (Joe Reis, 2022).

Data integration platforms offer a robust infrastructure that simplifies the intricacies of ETL. They provide the means to extract data from a myriad of sources, ranging from databases to cloud services, and subsequently transform this data into a standardized and analyzable format (Joe Reis, 2022). The transformation phase is vital, as it makes the data usable for a variety of applications, including analytics and machine learning. Moreover, these platforms ensure the seamless loading of processed data into the target systems, which can encompass data warehouses, databases, or other analytics tools. This strategic approach sets the stage for informed decision-making, by rendering data insights accessible to business analysts, data scientists, and other stakeholders (Kimball, 2019).

In the broader context, the seamless integration of data platforms not only streamlines ETL processes but also establishes a foundation for advanced analytics, including machine learning. As organizations extract, transform, and load data through these platforms, the standardized and enriched datasets become instrumental for conducting sophisticated machine learning analyses (Kimball, 2019). Specifically, when it comes to predicting customer churn, this integrated approach ensures that the data required for machine learning models is prepared, consistent, and readily available.

The transformation phase within the data integration platforms plays a crucial role in shaping the data into a format suitable for training predictive models. This enriched data, free from inconsistencies, serves as the fuel for machine learning algorithms, allowing organizations to derive meaningful insights into customer behavior and make informed predictions about potential churn. Therefore, the utilization of data integration platforms is not just about optimizing ETL processes; it becomes a strategic enabler for organizations to seamlessly transition from data processing to advanced analytics, ultimately enhancing their capability to predict and address customer churn effectively.

## Related works

We looked at some of the technologies and practices that have been put in place to tackle the need for data integration

### Machine Learning for Environmental Data Integration in Conservation Initiatives

Machine Learning for Environmental Data Integration in Conservation Initiatives, delves into the Santos Project's approach to addressing challenges in integrating environmental data for conservation and monitoring. The project proposed the use of machine learning techniques, specifically Random Forest (RF) for univariate data and a hybrid strategy combining Self-Organizing Map (SOM) and RF for multivariate data (Fonseca, 2023). They address challenges such as non-linearity, covariation, interactive effects, missing or noisy data, model optimization, accurate predictions, and imbalanced observations.

It emphasizes the critical need for conservation efforts in the face of environmental challenges. The project highlights the complexities of integrating diverse environmental data and outline the analytical challenges, which are contextualized within the Santos Project aimed at understanding the spatio-temporal dynamics of benthic, pelagic, and physical systems (Fonseca, 2023). The machine learning techniques highlighted in the project, RF for univariate data and a hybrid Self Organizing Map SOM - Random Forest (RF) strategy for multivariate data, are justified based on their ability to handle non-linearity, covariation, and interactive effects (Fonseca, 2023). The project emphasizes the advantages of RF, such as its ability to deal with non-linear data and variable importance analysis. For multivariate data, the combination of SOM and RF is presented as an effective approach.

Project touches the importance of addressing missing values, highlighting the superiority of the bagging imputation technique in handling missing or noisy data. The resilience of the machine learning techniques against noisy data is noted, and the identification of noisy data based on model outputs is highlighted. In the context of imbalanced data sets, the project explores the correlation between the overall statistics of the RF model and those of individual classes. This joint interpretation is deemed valuable for understanding model limitations and discussing environmental mechanisms shaping observed patterns.

They showed two analytical workflows that enable the exploration and enhancement of model accuracy while facilitating the investigation of cause-and-effect relationships in the data. These workflows are positioned as foundational for implementing long-term learning algorithms, crucial for monitoring initiatives. This is one of the cases we see how data integration is vital when it comes to analytical processing of data and machine learning

### DataOps

This is a new innovative approach to data integration that emphasizes on collaboration, automation and continuous delivery in the context of data-related processes (Munappy, 2020). It is a set of principles that aim to improve the collaborating and communication between data engineers, data scientists and other data professionals while also enhancing the efficiently, quality and agility of data related processes. It draws its inspiration from DevOps which its focus is improving software development and IT operations.

Some of the key elements of DataOps include close collaboration between various teams involved in data such as data engineers and data scientists to foster better communication and alignment of goals. Automation is another aspect that involves the automation of data related tasks such as data ingesting transformation and deployment to minimize manual effort and reduce chances of errors (Munappy, 2020). The other is continuous delivers in promoting the concept of continuous delivery of data, which means making data available for analytics, reporting and decision making in real-time or near real-time. Lastly is monitoring of feedback that is done for data pipelines, workflows and data quality and providing timely feedback for continuous improvement (Munappy, 2020).

### Customer churn prediction in telecom using machine learning in big data platform

The project addressed the significant issue of customer churn, particularly in the tele-communications industry, where retaining customers is crucial for revenue stability. The aim was to develop an efficient churn prediction model using machine learning techniques on a big data platform, focusing on feature engineering and selection. The research utilized a vast dataset from SyriaTel Telecom Company, spanning nine months and comprising structured, semi-structured, and unstructured data formats, totaling about 70 Terabytes on the Hadoop Distributed File System (HDFS). The project leveraged the Apache Spark environment to experiment with four machine learning algorithms: Decision Tree, Random Forest, Gradient Boosted Machine Tree (GBM), and Extreme Gradient Boosting (XGBOOST). The XGBOOST algorithm yielded the best results for churn prediction classification.

One notable contribution of the project was the incorporation of Social Network Analysis (SNA) features extracted from customer social networks into the prediction model. The SNA features enhanced the model's performance significantly, improving the Area Under Curve (AUC) standard measure from 84% to 93.3%. The dataset's challenges included its massive size, varied data formats, and rapid influx, necessitating a suitable big data platform for handling and processing. The project showcased the advantages of big data platforms in collecting, storing, and processing diverse data efficiently. The study also emphasized the importance of addressing class imbalance in the dataset, with three scenarios explored: oversampling, under sampling, and without re-balancing. The evaluation utilized AUC due to its suitability for unbalanced datasets. The research highlighted the superiority of big data systems in handling complex computations, particularly in the context of Social Network Analysis measures on large-scale networks.

The project's relevance was underscored by the telecom industry's focus on customer retention strategies, acknowledging the higher costs associated with acquiring new customers compared to retaining existing ones. The project demonstrated that big data technologies facilitated feature engineering, a critical aspect of predictive model development, allowing for the extraction of richer features. The model's evaluation and impact testing were conducted using a new dataset, and the successful deployment of the model to production further validated its effectiveness. The project provided an overview of existing studies in churn prediction, emphasizing the project's unique contribution in the Syrian telecommunications context. The research bridged gaps by addressing feature engineering from raw data, handling class imbalance, and utilizing big data technologies for enhanced predictive modeling. The project's methodology and findings contribute valuable insights to the field of churn prediction in the telecommunications sector.

## Gaps Customer churn with data integration

In the case of customer churn prediction systems with integrated data, there are notable opportunities for improvement. Currently, existing models predominantly concentrate on batch data processing, introducing a delay in obtaining insights and taking proactive measures. Recognizing this gap, our objective is to introduce real-time data integration capabilities. By facilitating the seamless integration of data sources in real-time, even on a small scale, we aim to enhance the system's agility and responsiveness to evolving customer behaviors. This shift toward real-time processing represents a crucial advancement, ensuring that businesses can promptly identify and address factors contributing to customer churn.

Furthermore, scalability and efficiency in data integration processes are essential for the seamless operation of predictive models. We recognize the need to explore solutions that not only enhance the scalability of the system but also improve its overall efficiency. This involves optimizing data integration workflows to handle growing datasets without compromising on performance. A scalable and efficient data integration process is fundamental to ensuring that the churn prediction system can accommodate increasing volumes of data while maintaining a high level of accuracy and responsiveness.

Ethical considerations are paramount, particularly in the realm of customer churn prediction where fairness, bias, and data privacy play crucial roles. As part of our commitment to ethical practices, we aim to address these concerns comprehensively. In churn prediction, biases may inadvertently impact certain customer segments, and the potential for privacy breaches is a serious consideration. Our approach involves incorporating ethical safeguards into the system, mitigating biases, ensuring fairness, and prioritizing data privacy. By proactively addressing these ethical considerations, we aim to build a customer churn prediction system that not only delivers accurate insights but also upholds the highest standards of fairness and privacy in its operations.

## Conceptual Frame work

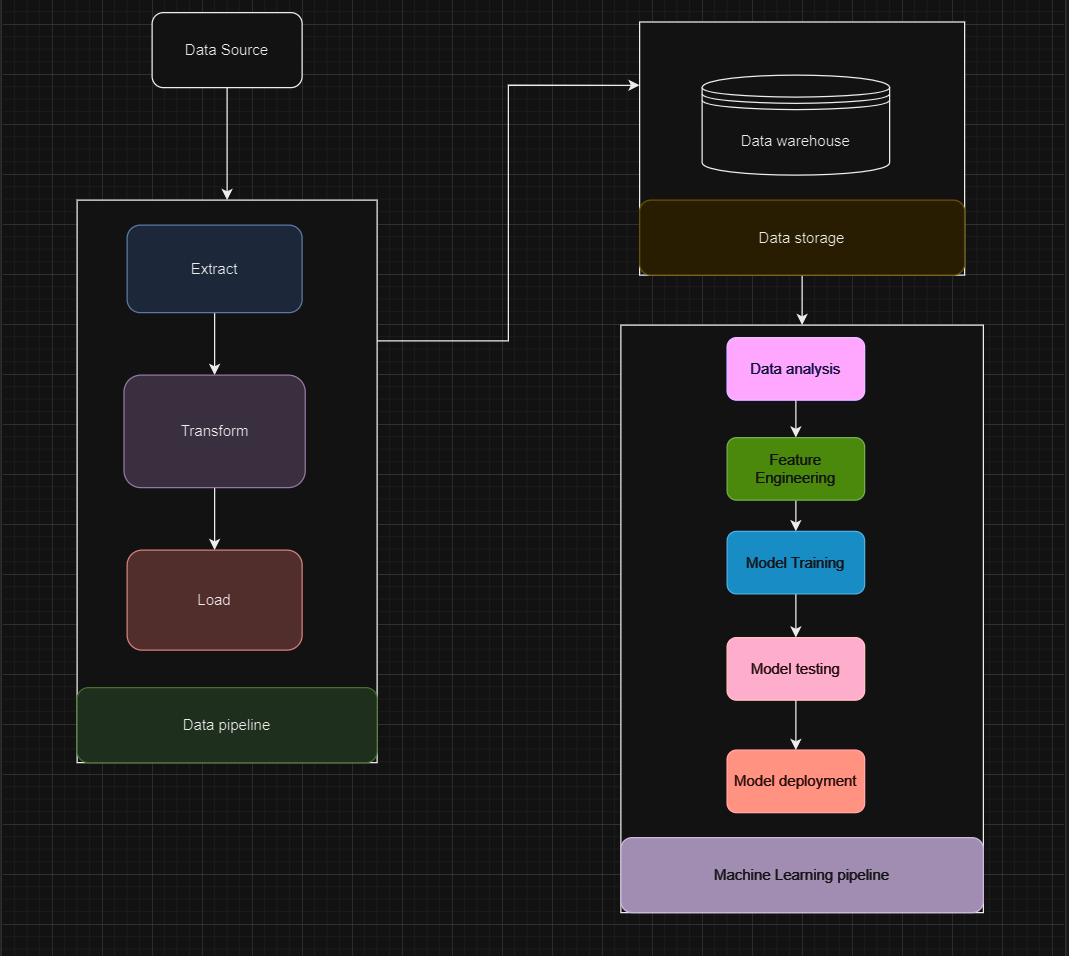


Figure 2.2 Conceptual Frame work

Above we have the conceptual frame work of how the project is going to look like. Let us see what each step represents and what happened.

### Data source

Data source is where the data source was be located, it could be in a database, retrieved via API or even sensors for cases in where the situation is big data. This data can be in various forms so some work needs to be done to it to fit the goal that is set for the business and in this case would be catered towards churning. This data was loaded into our data pipeline

### Extraction

This is the stem where the data is brought into the data pipeline so that various operation can be done to it. The extraction process depended on the nature of the data and where it is being retrieved from. In our case the data was retrieved from a csv file. So, the extraction was done to cater for csv files. The file was located in the same system as our data pipeline

### Transformation

The transformation stage is almost similar with data cleaning and processing. The raw data may contain errors or inconsistencies. This is where we apply a series of cleaning techniques to rectify the issues for example applying statistical methods and imputation of missing values and data inconsistencies may also be resolved. This stage ensured data is accurate and lack irregularities setting a good foundation for subsequent analysis (Bhatia, 2019).

For scenarios requiring data summarization or aggregation for reporting and analysis, the "Transformation" stage is the hub for such operations. Aggregation involves grouping and summarizing data to glean higher-level insights, employing operations like averaging, summing, or identifying maximum and minimum values within specific time periods or categories. This yields aggregated data offering a more concise and comprehensive view of trends and patterns within the dataset.

For this project under the data pipeline, it was supported by Apache Airflow which its main work is just to present a clear visual of how our data pipeline is behaving. The data pipeline eventually breaks down and this helped us to quickly and swiftly debug the program.

### Load

For this step the data had already undergone the necessary transformation that are in line with our business needs. This data now needs to be sent to a storage system in its current state. The data was loaded into a data warehouse. This was because it was geared towards analyzing it hence it cannot be used in the same database that is used by the business as it is optimized for online transactional process (OLTP) but for the storage we are using for our project was optimized for online analytical processes (OLAP) since the data was used for data analytics and machine learning

### Machine learning pipeline

Machine learning is where the steps that are happening in our machine learning pipeline and explain them.

Data analysis was the first step in the machine learning pipeline. In this step, we examine the data we have in detail. We aimed to understand its features, such as the type of information it holds, the volume of the data, and if there are any missing or unusual values. This step helped us understand the data better and guides us on how to prepare it for the next steps. The next step was featuring engineering. In this step, we prepared our data for the machine learning model. We needed to combine some data, divide other data, even create new data that assisted the model in making better predictions.

After our data was prepared, we proceeded to model training. In this step, we fed our prepared data to the machine learning model. The model attempted to learn patterns in the data. Initially, it did not perform very well, but with more training, it improves at identifying patterns. Once the model is trained, we needed to test it. In model testing, we presented the model with new data that it hasn't seen before. We then check how well the model's predictions align with the actual values. This gives us an idea of how well the model had learned from the training data and how well it can generalize to new data.

The final step was model deployment. Once the model was trained and tested, it's ready to be used on real-world data to make predictions. All this happened within Jupyter notebook which is a popular IDE used for machine learning in python

### Underlying Factors

#### Security

In the realm of e-commerce, security is paramount. Ensuring the security of customer data is not only a legal requirement but also a trust-building measure. This involves measures like encrypting customer data during storage and transmission, implementing access controls to restrict data access to authorized personnel only, and monitoring for any unusual or unauthorized activities that may indicate a security breach. Compliance with data protection regulations, such as the GDPR, is essential to maintain customer trust and avoid legal issues (Ralph Kimball, 2019).

#### Data Management

E-commerce platforms generate a vast amount of data daily. Effective data management is essential for organizing and maintaining this data. It includes structuring data storage to handle the large volume of transactional data, implementing indexing mechanisms for quick data retrieval, version control to track changes and ensure data integrity, and data cataloging to ensure that data is well-organized and easily accessible. Proper data management helps in efficiently handling and using the wealth of customer data generated (Joe Reis, 2022).

#### DataOps

In the e-commerce context, DataOps can significantly improve the efficiency of the data pipeline. It combines data engineering and operations to streamline data processes. Collaboration among data engineers, data scientists, and operations teams is vital to ensure data flows seamlessly. Automation of routine data tasks, monitoring for data quality, and rapid deployment of updated models are essential aspects of DataOps. For an e-commerce platform, DataOps can lead to timely insights into customer behavior, allowing for swift responses to potential churn risks (Joe Reis, 2022).

#### Data Architecture

Data architecture plays a critical role in an e-commerce platform's ability to efficiently store and manage data. Decisions about databases, data models, and the overall structure of the data ecosystem are key. In e-commerce, data architecture needs to consider the scalability and performance requirements of handling vast quantities of transactional data. Ensuring that the architecture supports both real-time and historical data is crucial for effective customer churn prediction (Joe Reis, 2022).

#### Orchestration

Effective orchestration is vital for coordinating data processes and tasks in the pipeline. In e-commerce, it's about managing the flow of data from different sources, including online transactions, customer interactions, and other data streams. Workflow automation ensures that data is processed efficiently and in the right sequence. For example, orchestrating the flow of data from website clicks to purchase history can provide a comprehensive view of customer behavior (Joe Reis, 2022).

#### Software Engineering

Software engineering principles are central to the development and maintenance of the software components of our data pipeline. This includes building APIs for data access, creating data transformation scripts to preprocess and clean data, and developing machine learning models for customer churn prediction. In e-commerce, efficient software engineering ensures that the churn prediction models are up-to-date, scalable, and responsive to real-time customer interactions (Joe Reis, 2022).

# Methodology

## Introduction

In this chapter we looked at the methodology that is going to be selected for our project and in this case is the building a data pipeline for our customer churn prediction in the e-commerce sector. We looked at overview of the development approach that we had select and deep dive in to the different phases of the project. The chapter entails the introduction to the selected development approach, a justification for its selection and in-depth system analysis and a thorough system design.

## Development Approach

For our project, we carefully considered that the selection of a development approach that is well suited for the complexities in addressing data integration for a customer churn prediction in the ecommerce industry. And after evaluation we chose to go with the structured systems analysis design methodology. SSAD is the systematic approach that provided a well-defined framework for analyzing, designing and implementing complex systems also offering a structured and organized way to tackle intricate projects making it valid choice for our data engineering project

## Justification of the Methodology

The choice for the methodology that we chose which is the SSAD is because of its effectiveness in handling complex projects especially in this case where we are looking to integrate data into a customer churn prediction in the e-commerce sector. SSAD provides clear and organized approach to system development, which is crucial for a project with multiple components and data integration requirements. It emphasized the importance of thorough analysis, logical design and efficient implementation ensuring that the project aligns with the specific needs and objectives of the e-commerce business

## System Analysis

In this phase, we utilized various SSAD tools and techniques to comprehensively understand the requirements of our project and build our system in accordance. The focuses were on defining the technologies, assessing team workloads, identifying limitations, establishing timeframes, and managing the budget. This methodology provided a structured and systemic approach to system analysis, which was invaluable in ensuring the success of our data engineering project

### Context Diagram

A context diagram is a simplified visual representation used in data analytics to illustrate interactions between a data pipeline and its external environment. In the context of a data analytics project for an e-commerce company, the system (data pipeline) is depicted along with external entities like the online store database, customer data, and a management dashboard. Processes, such as data cleaning and sales analysis, are represented within the system, connected by arrows indicating the flow of data.

The significance of a context diagram was its ability to provide a clear and concise overview of the system's interactions. It helped define the project's scope, facilitated communication among team members and stakeholders, and served as a foundational tool for understanding the flow of data within our system.

### DFD Level 0, 1

In the context of our project, the Level 0 Data Flow Diagram (DFD) served as a strategic starting point, offering a bird's-eye view of the entire system. Here, a single process encapsulated the entirety of our data analytics pipeline, connecting with external entities such as databases, users, and potentially a reporting system. The Level 0 DFD laid out the fundamental processes, key data stores, and the essential data flows that illustrate how information was exchanged between the system and its external elements.

As we progressed to the Level 1 Data Flow Diagram (DFD), the diagram provided a more detailed breakdown of the major processes identified at Level 0. Each primary process, such as data cleaning or sales analysis, is further dissected into specific sub processes. This decomposition created a more granular perspective, highlighting the intricacies of the internal workings of our data integration system. Additionally, Level 1 incorporates data stores, representing where data is stored during various stages, and depicts the data flow paths, outlining precisely how data moves between sub processes and stores. This level of detail was instrumental for a comprehensive understanding of the system's internal processes and the flow of data within them.

### ERD (Entity-Relationship Diagram)

In our data engineering project, the Entity-Relationship Diagram (ERD) served as a valuable tool for modeling the relationships between entities within our data pipeline. Entities, representing distinct elements in our system like datasets or tables, are visualized along with their attributes, offering a comprehensive view of the data's characteristics. For instance, in an e-commerce analytics scenario, entities might include "Customer" and "Product," each with attributes defining customer details and product information.

The ERD's primary function is to illustrate how these entities are interrelated. Relationships, depicted by lines connecting entities, showcase how data flows and interacts within the system. This visual representation aids in understanding the data's structure and organization, providing a foundation for designing an efficient and cohesive data analytics pipeline.

### Database Schema

In our data engineering project, the Database Schema acted as a crucial blueprint, detailing the structure and organization of our database. It outlines tables such as "Customers," "Products," and "Sales," each with specific fields like customer ID, product ID, and sales data. The schema not only defines the data types and attributes within these tables but also establishes relationships between them, providing a clear roadmap for how data is stored and interrelated.

The significance of the Database Schema lies in its ability to offer structural clarity and define relationships between different elements of data. For instance, the schema specifies how customer information in the "Customers" table is linked to sales data in the "Sales" table through unique identifiers. This clear representation facilitates efficient data retrieval and storage, forming a foundational element in the design and maintenance of a robust data engineering pipeline.

## System Design

In the system design phase of our data engineering project aimed at creating an E-commerce Customer Churn Prediction system, we leveraged various Structured Systems Analysis and Design (SSAD) tools and techniques to ensure a comprehensive understanding of project requirements. The focus was on defining technologies, assessing team workloads, identifying limitations, establishing timeframes, and managing the budget. This systematic approach aims to lay a solid foundation for the success of our project.

The Context Diagram served as a crucial guide in visualizing the interactions between our data pipeline and external elements. Representing the system (data pipeline) alongside entities like the online store database, customer data, and a management dashboard, this diagram helped define the project's scope and facilitate effective communication among team members and stakeholders.

The Data Flow Diagrams (DFDs) at Levels 0 and 1 provided a structured overview of our data analytics pipeline. The Level 0 DFD offers a bird's-eye view, encapsulating the entire system and illustrating fundamental processes, data stores, and essential data flows. As we move to the Level 1 DFD, a more detailed breakdown of major processes, such as data cleaning and sales analysis, was presented. This granular perspective, including sub processes, data stores, and data flow paths, is crucial for a comprehensive understanding of the internal workings of our customer churn prediction system.

The Entity-Relationship Diagram (ERD) played a pivotal role in modeling relationships between entities within our data pipeline. Entities like "Customer" and "Product," each with specific attributes, was visualized, demonstrating the interrelations crucial for predicting customer churn. This visual representation aids in understanding the data's structure and organization, providing the foundational structure for designing an efficient and cohesive E-commerce Customer Churn Prediction system.

The Database Schema not only defines data types and attributes within tables but also specifies how different elements of data are connected. For instance, it outlines how customer information in the "Customers" table is linked to sales data in the "Sales" table through unique identifiers. The Database Schema's significance lies in providing structural clarity and facilitating efficient data retrieval and storage. This schema forms a foundational element in the design and maintenance of a robust data engineering pipeline, ensuring the seamless functioning of our E-commerce Customer Churn Prediction system.

## System Deliverable/Milestones

### Deliverables

The feasibility study report systematically evaluated technical feasibility by assessing the technology stack's compatibility with the project requirements. Operational feasibility delved into how well the proposed data pipeline aligns with existing business processes. Economic feasibility included a detailed cost-benefit analysis, considering both initial development costs and long-term operational expenses.

Requirements Specification Document not only listed functional and non-functional requirements but also detail the rationale behind each requirement. It provided a traceability matrix to link requirements to specific project objectives, ensuring transparency and alignment with the overarching goals of predicting customer churn in the e-commerce sector.

Building on the context diagram, the system architecture diagrams included an expanded DFD (Data Flow Diagram) hierarchy, illustrating the flow of data through different system components. Additionally, the ERD (Entity-Relationship Diagram) modeled the relationships between various entities in the e-commerce data environment, guiding the database schema design.

The implementation codebase for workflow orchestration and machine learning algorithm was organized following best practices and coding standards. Detailed comments and documentation within the code facilitated ease of understanding for future developers or maintainers. Version control systems like Git was used to track changes, ensuring code reliability and collaboration among team members.

The testing documentation included unit tests for individual components, integration tests to ensure seamless interaction between different modules, and system tests to validate the overall functionality. Performance testing was conducted to assess the system's response under various loads, ensuring its scalability.

User manuals catered to various stakeholders, providing different levels of detail for end-users, system administrators, and developers. They included step-by-step guides, screenshots, and troubleshooting sections to enhance usability and facilitate a smooth user experience.

### Milestones

The completion of the feasibility study involved presenting findings to stakeholders, engaging in discussions to address any concerns, and obtaining formal approval to proceed. It marks a critical point where the project's potential challenges and risks are well-understood.

The sign-off on requirements analysis involved stakeholders acknowledging that the specified requirements accurately reflect their needs. This milestone is crucial to ensure that subsequent development phases are aligned with the business objectives and expectations.

These phases were organized using an agile framework, with sprints and iterative development. Milestones may include the completion of specific user stories, successful integration of data sources, and the achievement of predefined testing criteria. Regular sprint reviews kept stakeholders informed and engaged.

Project completion involves deploying the data pipeline to the production environment, monitoring its initial performance, and transitioning to the maintenance and support phase. A comprehensive handover to the operations team ensured a smooth transition and ongoing stability.

### System proposal

The proposed Data-Driven Customer Churn Prediction System is a comprehensive solution designed to tackle the persistent challenge of customer churn faced by businesses. Leveraging sophisticated data integration techniques, the system seamlessly incorporates diverse data sources, offering a holistic view of customer interactions. By employing advanced machine learning algorithms, the system analyzes historical customer data to predict potential churn indicators, identifying patterns and correlations. Real-time monitoring ensures continuous tracking of customer behavior, enabling immediate responses to evolving patterns and potential churn triggers. The system presents its findings through an intuitive dashboard, providing stakeholders with actionable insights for informed decision-making. Expected outcomes include proactive churn mitigation, improved decision-making through data-driven insights, and enhanced customer retention rates. The implementation plan encompasses phases such as system integration, data onboarding, algorithm implementation, real-time monitoring setup, and dashboard development, ensuring a seamless and effective deployment of the proposed system into existing business operations.

# System Analysis and Design

## Introduction

This chapter delves into the analysis and design of the Data-Driven Customer Churn Prediction System, a comprehensive solution designed to tackle the persistent challenge of customer churn faced by businesses. The system leverages sophisticated data integration techniques and advanced machine learning algorithms to predict potential churn indicators, identifying patterns and correlations.

We explored the system requirements, both functional and non-functional, that underpin the system's design and operation. This included the data pipelines, the data exploration and machine learning methods that we used.

We also examined the various system analysis and design diagrams that have been drawn up to illustrate the system's structure and workflow. These include Context Diagram, Data Flow Diagrams, Activity Diagrams, ERD Diagrams, Logical Database Schema, and Class Diagrams.

The chapter also discusses the implementation plan, which encompasses phases such as system integration, data onboarding, algorithm implementation, real-time monitoring setup, and dashboard development. This plan ensures a seamless and effective deployment of the proposed system into existing business operations.

By the end of this chapter, the reader has a comprehensive understanding of the system's design, its requirements, and the steps taken to ensure its successful implementation and operation.

## System Requirements/Requirements Gathering

The Data-Driven Customer Churn Prediction System is designed with a set of specific functional and non-functional requirements in mind. These requirements are crucial to ensure the system's effectiveness in predicting customer churn and providing actionable insights for businesses.

### Functional Requirements

1. Data Integration: The system is designed to seamlessly incorporate diverse data sources, offering a holistic view of customer interactions. This is achieved through sophisticated data integration techniques in our case would be for example extracting data from various files such as csv.
2. Machine Learning Algorithm: The system employs advanced machine learning algorithms to analyze historical customer data. This allows the system to predict potential churn indicators by identifying patterns and correlations. Some of the algorithms we used were for example Random Forest and Decision Trees.
3. Real-Time Monitoring: The system is equipped with real-time monitoring capabilities to continuously track data flow within pipelines. This high data through put and high data quality and integrity.
4. Dashboard Development: The system presents its findings through an intuitive dashboard, providing stakeholders with actionable insights for informed decision-making.

### Non-Functional Requirements

1. System Integration: The system is designed to be integrated into existing business operations seamlessly. This is achieved through a detailed implementation plan that includes system integration, data onboarding, algorithm implementation, real-time monitoring setup, and dashboard development.
2. System Security: The system is designed with robust security measures to protect the data and the integrity of the system. This includes the use of secure data transfer protocols and stringent access controls.
3. Scalability: The system is designed to handle varying loads, ensuring its scalability. This is achieved through performance testing, which assesses the system's response under various loads.
4. Usability: The system is designed with an intuitive user interface and detailed user manuals. This ensures a smooth user experience and facilitates ease of understanding for various stakeholders.

## System Analysis Diagrams

System analysis diagrams provide a visual representation of the system's components and their interactions. They are crucial in understanding the system's functionality and identifying potential areas of improvement. In the context of the Data-Driven Customer Churn Prediction System, several diagrams were used to analyze the system.

### Context Diagrams

The context diagram provides a high-level overview of the system, illustrating the system's interactions with external entities. It shows how the system exchanges data with its environment, including users, data sources, and other systems. This is an overview example and wont dive into the most basic functions of the systems and was to just show from a level of high abstraction we expanded on it in higher level DFD diagrams

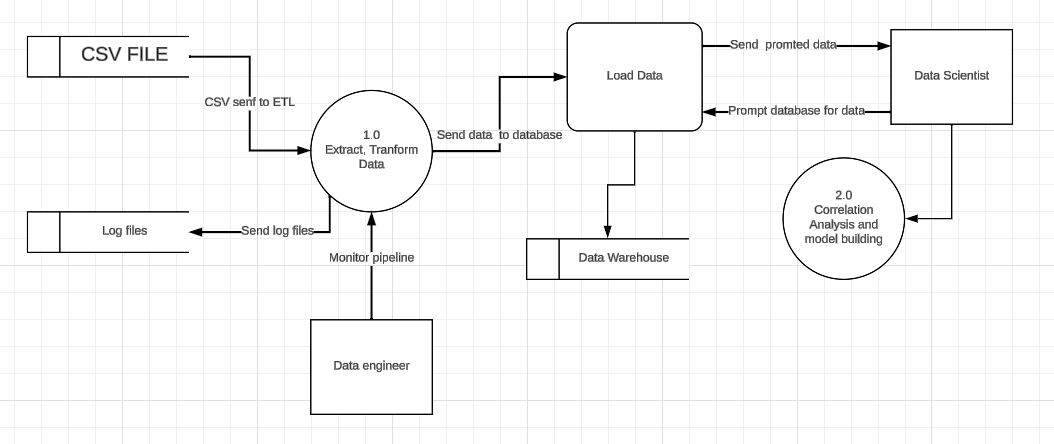


Figure 4.1 Context

### Data Flow Diagrams (Level 0, Level 1)

Data Flow Diagrams (DFDs) provide a detailed view of the system's data flows. The Level 0 DFD, also known as the context diagram, shows the system as a whole and how it interacts with external entities. The Level 1 DFD breaks down the system into its major high-level processes, illustrating how data flows between them. Shows how the diagrams starts with the files being loaded into the ETL pipeline and then to the database and forwarded to the data scientist for data exploration and model building and finally deployment. There was error handling here and there in case of failure in the data pipeline that was used by the data engineer who is in charge of managing the data pipeline.

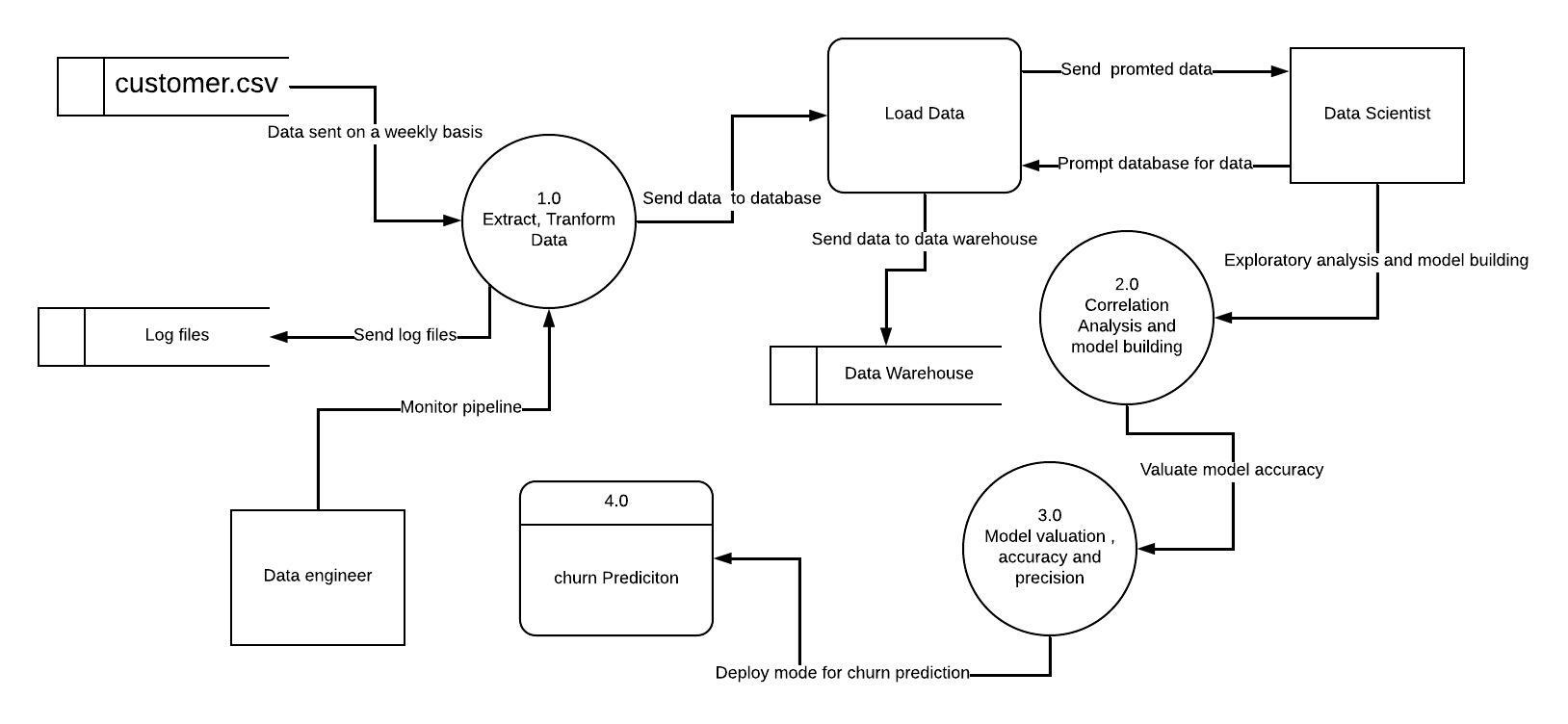


Figure 4.2 DFD level 1

### Activity Diagrams

Activity diagrams depict the system's dynamic behavior, showing the sequence of activities that occur in a particular use case. They are particularly useful in visualizing the flow of control in the system.

These diagrams were instrumental in understanding the system's data flows, interactions, and control flows. They provided valuable insights into the system's functionality and helped identify potential areas of improvement. The process starts with the files being loaded into the data pipeline from the csv files and begins the ETL process. We can see that depending on the error the data pipeline can either retry or fail and send a notification. After the etl process the data ended up in the database in the table structure that we defined in the ETL pipeline.

The data once is in the database it is ready for query from the data scientist for the machine learning process where he built and train model and finally deploy it for use. Our two main users here were the data engineer who dealt with the data pipeline by monitoring it in case of any failures by reading report logs and using airflow to see health of the pipeline. The other is the data scientist who dealt with the building deploying and training of the model, they worked hand in hand in order to maintain a healthy data ecosystem that was used in the business making decision.

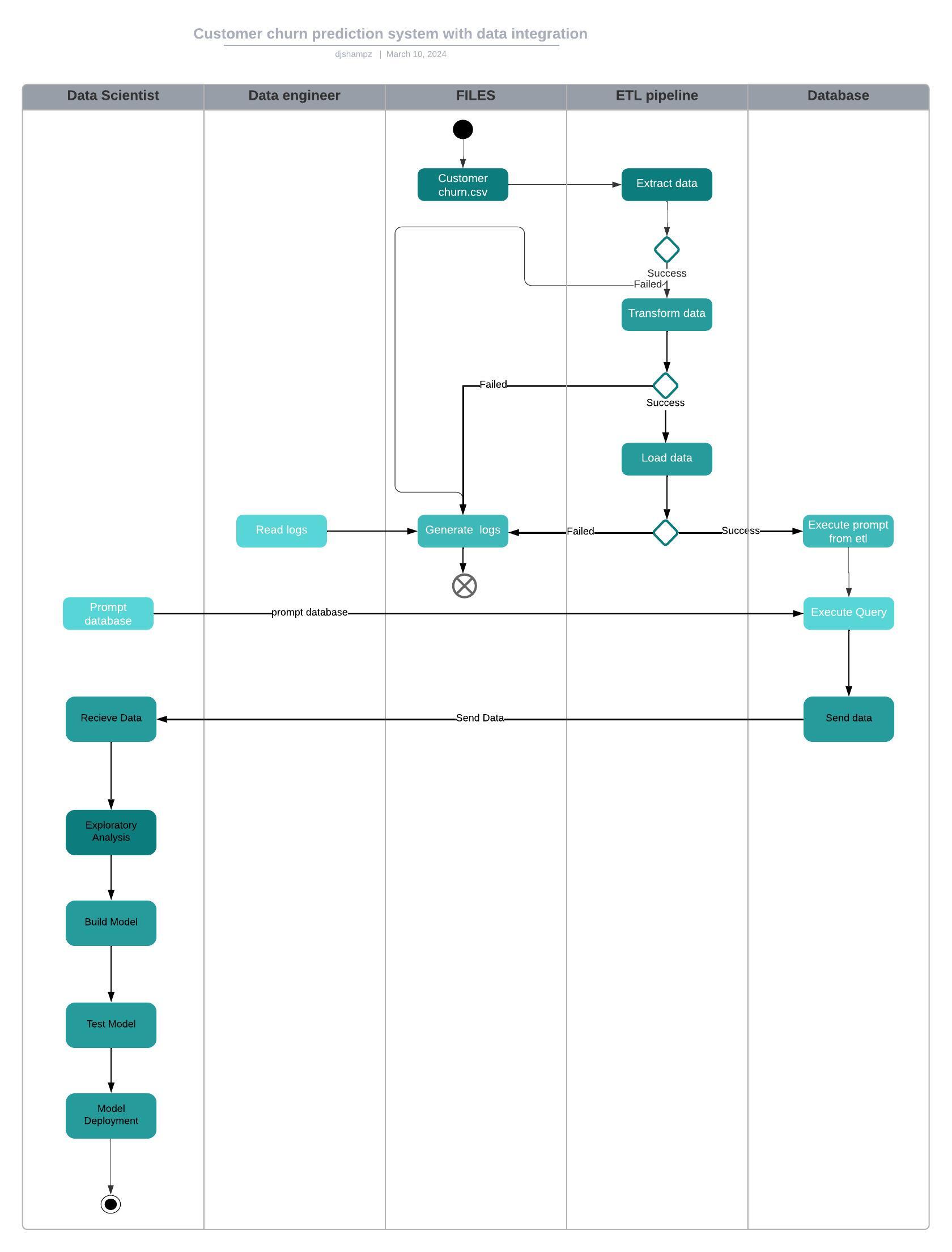


Figure 4.3 Flow chart

## System Design Diagrams

System design diagrams are instrumental in visualizing the structure of the system and how its components interact with each other. They provide a clear picture of the system's architecture, making it easier to understand and modify. In the context of the Data-Driven Customer Churn Prediction System, two types of diagrams were primarily used: Entity-Relationship Diagrams (ERD) and Logical Database Schema.

### Entity-Relationship Diagrams (ERD)

ERDs are used to model the relationships between various entities in the system. They provide a visual representation of the system's data and how it is organized, making it easier to understand the system's structure and the relationships between its components. In the context of the Data-Driven Customer Churn Prediction System, the ERD illustrates the relationships between various entities in the e-commerce data environment, guiding the database schema design.

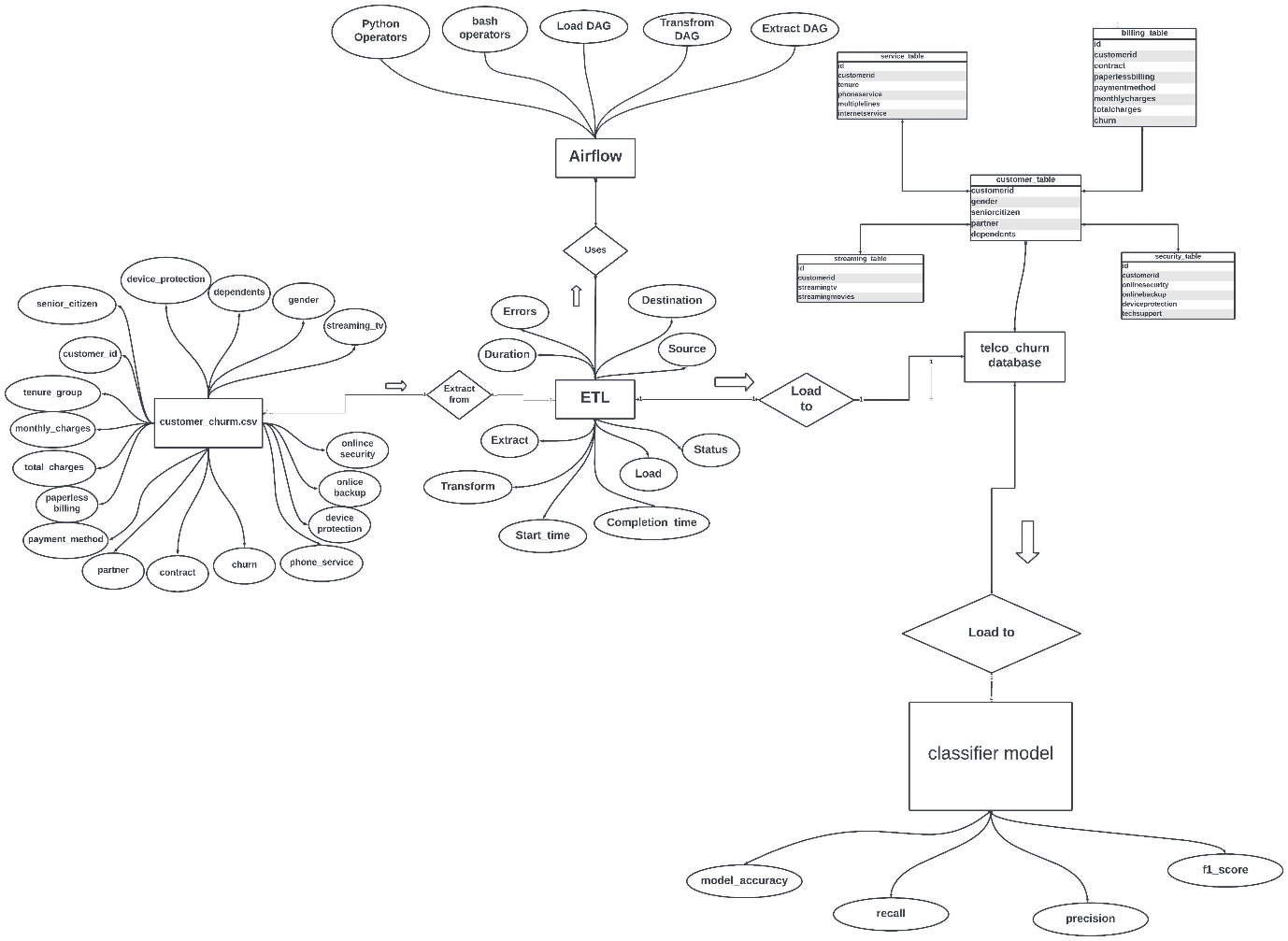


Figure 4.4ERD diagram

### Logical Database Schema

The Logical Database Schema shows how the actual database was formed by logically setting out how the tables and relationships were formed. It provides a detailed view of the system's data structure, including tables, fields, and the relationships between them. In the context of the Data-Driven Customer Churn Prediction System, the Logical Database Schema illustrates how customer data is organized and how it can be accessed and manipulated to predict customer churn.

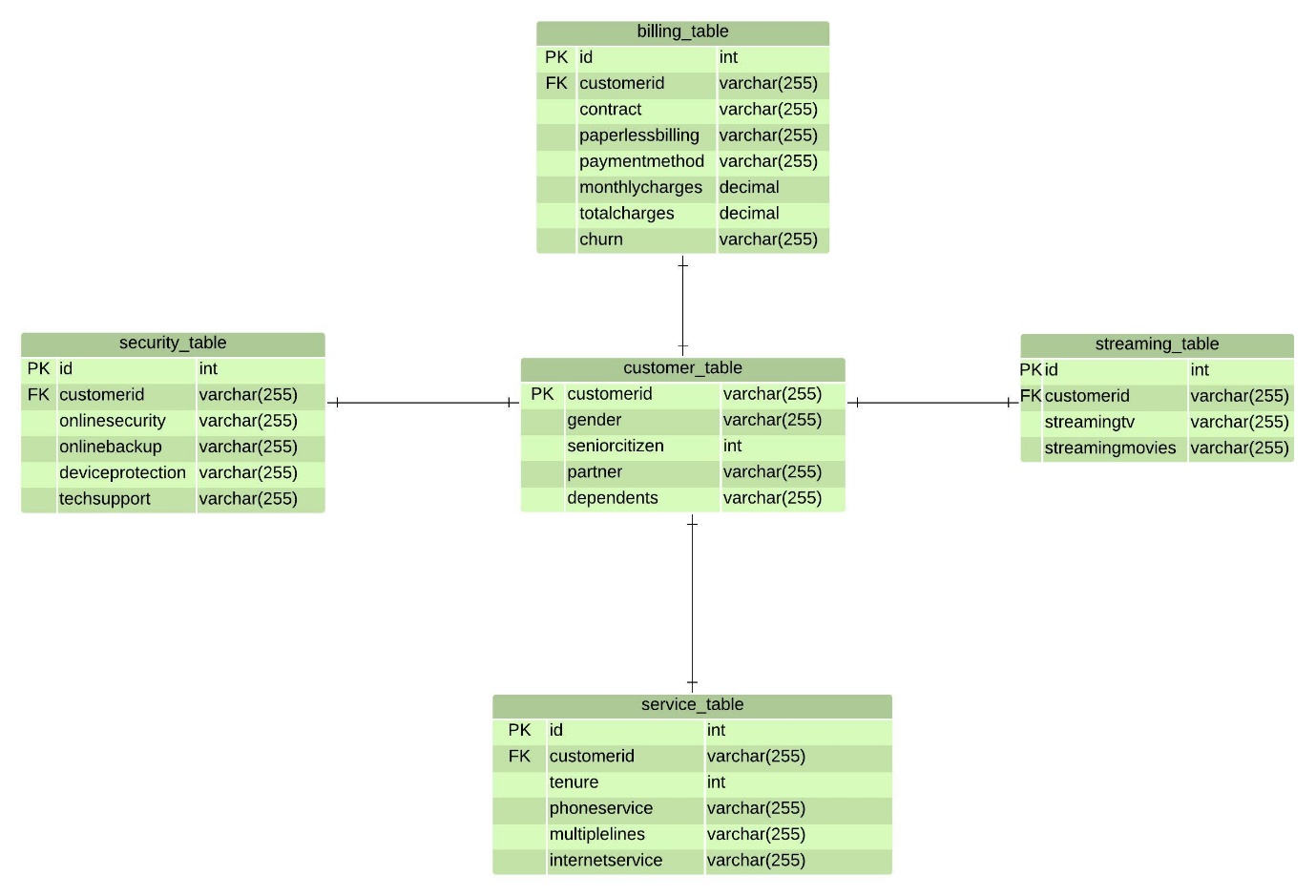


Figure 4.5 Database schema

These diagrams were instrumental in designing the system's data structure and understanding how data flows within the system. They provided valuable insights into the system's design and helped guide the implementation of the Data-Driven Customer Churn Prediction System.

# System Implementation and testing

## Introduction

This chapter delves into the practical aspects of the Data-Driven Customer Churn Prediction System, focusing on its implementation and testing. The implementation environment, both hardware and software, is described in detail, providing an understanding of the infrastructure that supports the system.

For machine learning projects, a comprehensive description of the dataset used is provided, including its retrieval, analysis, and division into training, testing, and validation data. The chapter also outlines the testing strategies employed to ensure the system's robustness and reliability. It discusses the various testing paradigms used, such as white box testing, black box testing, and unit testing, and presents the results of these tests.

The aim of this chapter is to provide a comprehensive understanding of how the system was brought to life from the design and analysis stages, and how its functionality and performance were validated. This provided insights into the practical challenges encountered and how they were addressed, thereby offering a complete picture of the system's development lifecycle.

## Description of the Implementation Environment

### Hardware Specifications

The development of the Data-Driven Customer Churn Prediction System was carried out on a laptop. The laptop's specifications are as follows:

* 1. The type of processor used during the development and testing of this system is the 13th Gen Intel(R) Core (TM) i5-13500X 20(CPUs) with 2.5GHz
  2. The amount of RAM used when running the system during implementation and testing is about 16GB
  3. The type of machine used for the running of the system during implementation and testing is Hp Omen 16.

The laptop provided a reliable and efficient environment for the development and testing of the system. Its processing power and memory capacity were sufficient to handle the computational demands of the machine learning algorithms and data processing tasks. The operating system supported all the necessary software tools and provided a user-friendly interface for development.

### Software Specifications

The software tools used for the development of the system are as follows:

* The database management system used for the storage of data entered in the system and any alterations made to the data during implementation and testing, was MySQL version 5.0
* The operating system used for the running of the system during implementation and testing was Ubuntu 20.0.1 which is 64-bit.
* The type of Web server used to host the system during implementation and testing was Apache version 3.2.2
* The web browser used during implementation and testing is Google Chrome.

These software tools were chosen for their robustness, efficiency, and compatibility with the project requirements. They provided a comprehensive environment for the implementation of the system, from data processing and machine learning to database management.

## Description of the Dataset

The dataset used for the development of the Data-Driven Customer Churn Prediction System is a CSV file containing customer data. The data was retrieved from the file WA\_Fn-UseC\_-Telco-Customer-Churn.csv and includes various customer attributes such as gender, senior citizen status, partner status, dependents, tenure, phone service, multiple lines, internet service, online security, online backup, device protection, tech support, streaming TV, streaming movies, contract, paperless billing, payment method, monthly charges, total charges, and churn status.

The dataset was analyzed using Python, with the help of pandas and NumPy libraries. The data was divided into training, testing, and validation sets to ensure a robust machine learning model. The features used for the model were extracted from the dataset, and one-hot encoding was applied to convert categorical data into a format that could be provided to the machine learning algorithms.

The dataset was processed using an ETL (Extract, Transform, Load) pipeline, as described in the etl.py script. The script includes functions for creating a database connection, extracting column names, creating column types, creating tables, inserting data into tables, and adding primary and foreign keys. The script also includes data cleaning steps such as removing duplicates, replacing empty cells with NaN, and converting the 'totalcharges' column to numeric. The dataset provided a comprehensive view of customer interactions, which was crucial for the prediction of customer churn. The features in the dataset allowed the machine learning model to identify patterns and correlations that could indicate potential churn. The use of this dataset was integral to the successful implementation of the Data-Driven Customer Churn Prediction System.

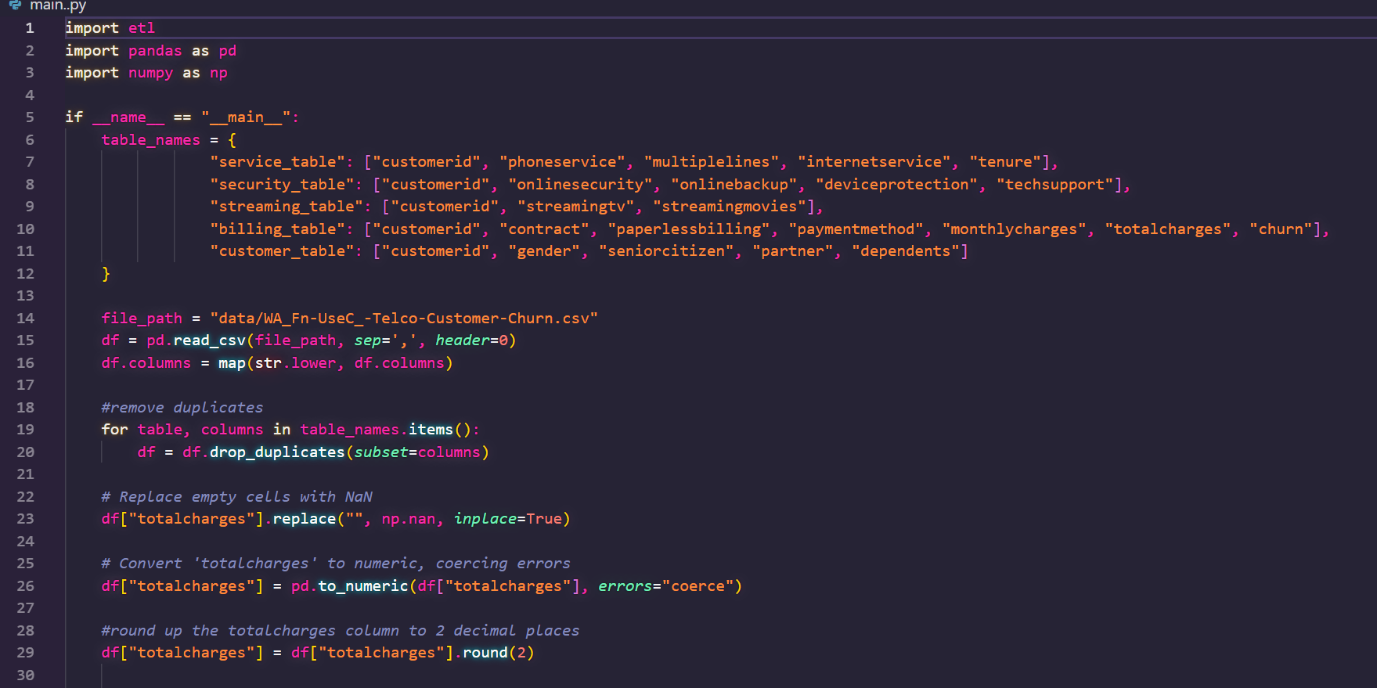


Figure 5.1 ETL code

The above screen shots just show some of the operation that were conducted on the data this happened within our data pipeline that extracted the data from a csv file named WA\_Fn-UseC\_-Telco-Customer-Churn.csv. The above shows a dictionary with lists of table columns for values and keys are the table names, that was used for create our table structure that was used as a basis of creating our database schema

All this process were executed in one of the dags in our airflow orchestration. There were three stages of our data pipeline which were

1. Extraction – this was basically be the extraction of data from the source destination and there won’t be much going on here as for our case the file was made available locally, there could be cases where data would be from APIs and cloud databases which our system is well equipped to do so
2. Transformation of the data – this entailed the transformation of data from the csv files into data frames for further processing. There were conversions of our columns from capital or uppercase to lowercase for consistency purposes when dealing with the column names. Also, some of the columns need to be converted into their proper datatypes
3. Loading – This stage loaded data into our MySQL server into their respective tables also creating various keys and references / relationships that were needed when dealing with the database and queries

## Airflow and work orchestration

As we have seen there are three jobs that we used in our pipeline. This was the ETL process which is the extraction, transforming and loading of data. All this was handled by airflow which is the tool that was used for the workflow orchestration. This is where the data engineer scheduled how often the jobs happened and when they happened. Also, here we were be able to read logs of the jobs processes. Airflow is coded in python and is where we injected our functions that we used to handle the data.

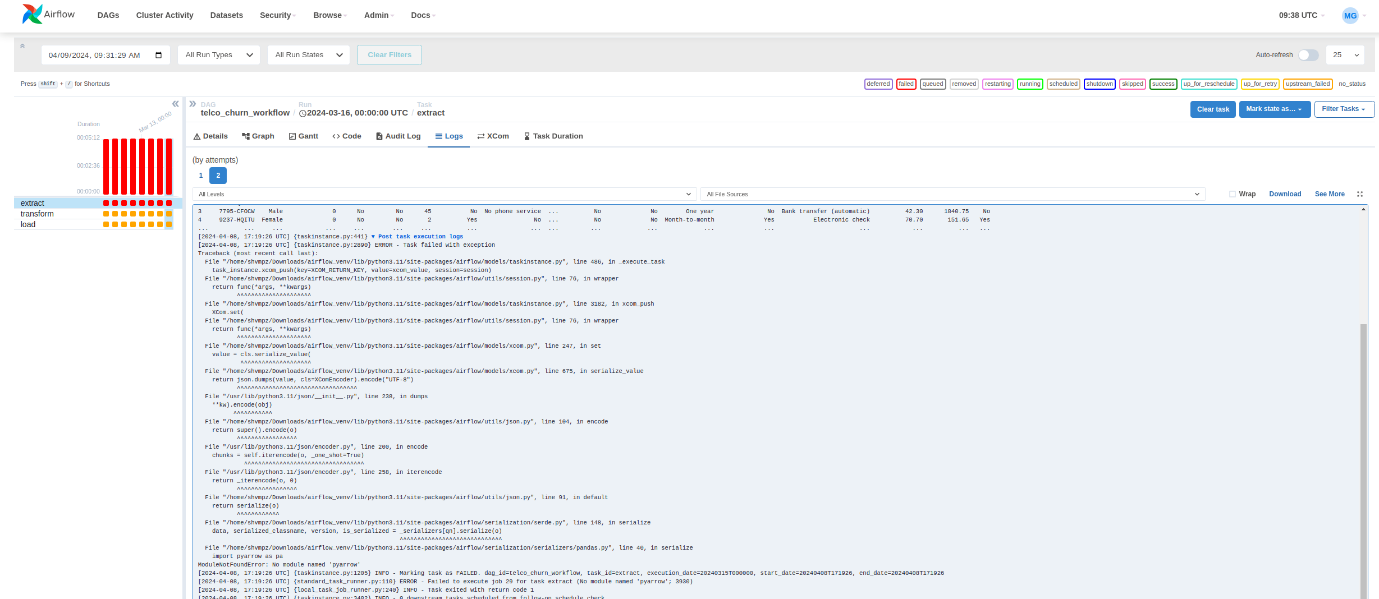


Figure 5.2 Airflow interface

This page of airflow highlights the scheduled job called telco\_churn which was the name of our DAG. The figure also highlights the error in our dag that caused it to fail and we can see there was a missing package not installed. This interface was ideal in trouble shooting and handling the health of the pipeline.

## Description of Testing

This section looks at the various ways of testing we used for our pipeline in general and our machine learning model. This entailed the various test case scenarios of how the pipe line should behave when put under certain situations and stress. For the model it would involve using various testing methods on how accuracy, f1-score and other metrics that are looked at in this chapter.

For the machine model we are focused on whether a customer churned or not so the churn column was be used as the goal of our project and the rest of the columns were used as the feature column that we used to show how they relate a customer churning.

### Exploratory data analysis

First and foremost, we conducted exploratory data analysis on our data to check on various aspects and relationships that our data has. We started by checking the number of churned and non-churned customers

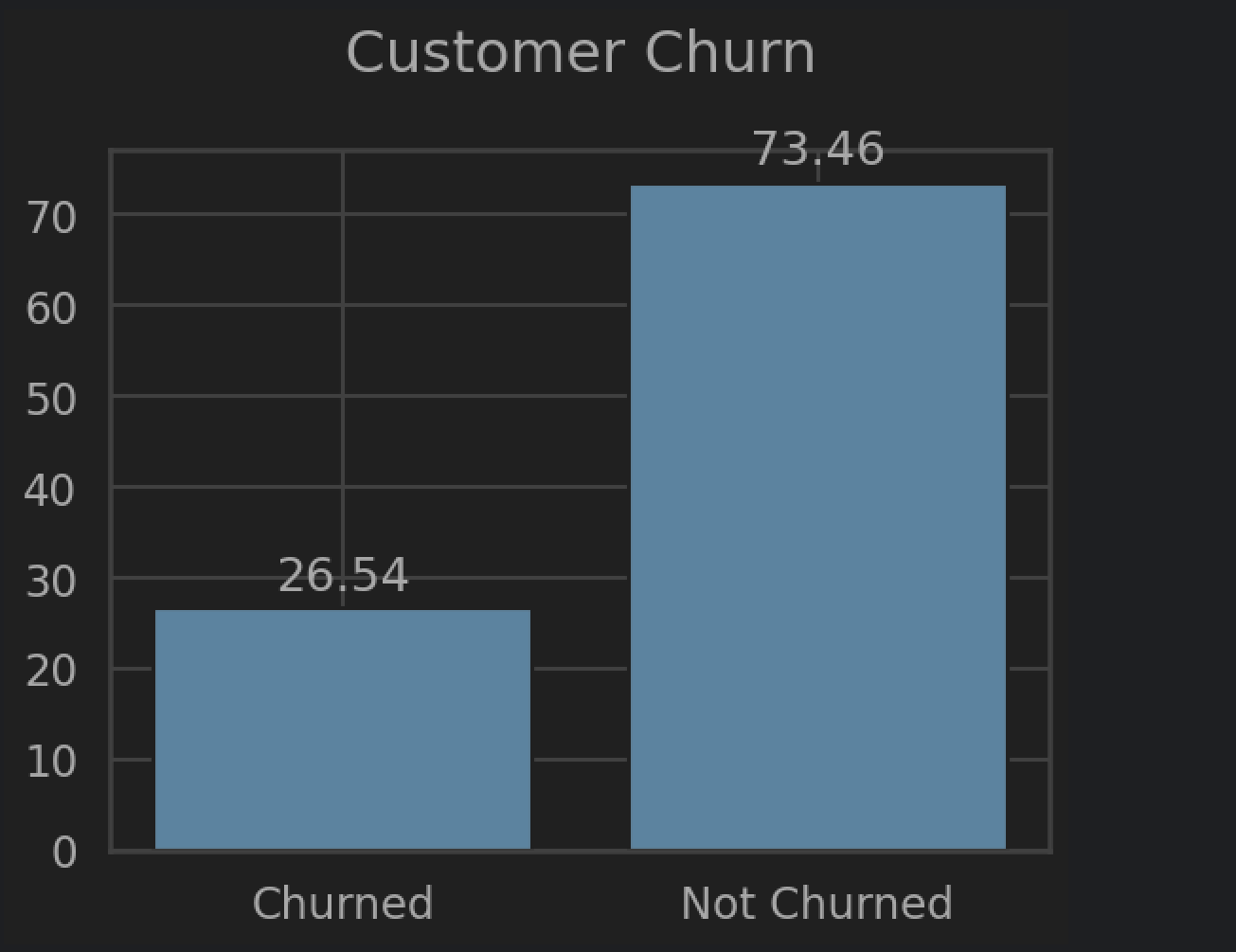


Figure 5.3 Ratio of churned to non-churned customers

From Figure 5.1, it's evident that our dataset is imbalanced, with a higher representation of non-churned customers. This imbalance is a critical factor to consider during our machine learning and model building process, as it can significantly impact the performance and reliability of our model. In such scenarios, accuracy, which is a commonly used metric, may not be the most reliable measure of model performance.

This is because accuracy calculates the proportion of all true predictions (both positive and negative) in the overall data. In an imbalanced dataset, a model that always predicts the majority class still achieved a high accuracy rate, even though it failed to correctly predict the minority class, which is often the class of interest. For instance, if we have a dataset where 95% of the instances belong to Class A (churned) and only 5% belong to Class B (non-churned), a model that always predicts Class A, regardless of the input features, would be 95% accurate. However, this model would be completely useless for predicting Class B, which could be the class of interest.

Therefore, in this context of imbalanced datasets, it's more appropriate to use other metrics such as precision, recall, F1 score, or Area Under the Receiver Operating Characteristic Curve (AUC-ROC), which take into account both false positives and false negatives, and are more sensitive to the performance of the model on the minority class. These metrics provided a more holistic view of the model's performance, ensuring that our model is not only accurate but also reliable.

Next, we look at how each feature and the its statistics relating to churning we looked at all columns but we listed just a few as examples.

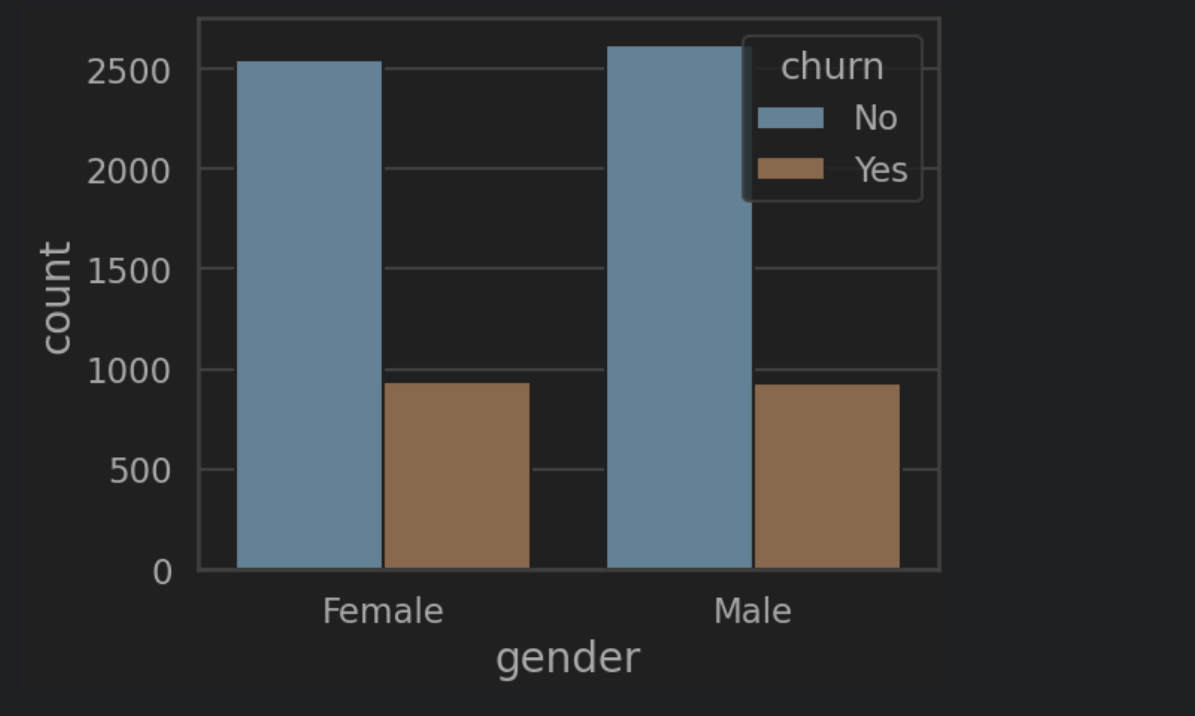


Figure 5.4 Female and male churning

From the above we see that churning is quite balanced when it comes to gender and there is not clear indication whether female or males have a higher chance of churning or not. So, in short, the graph suggests that while gender might be included as a feature, it may not significantly impact the model’s ability to predict churn.

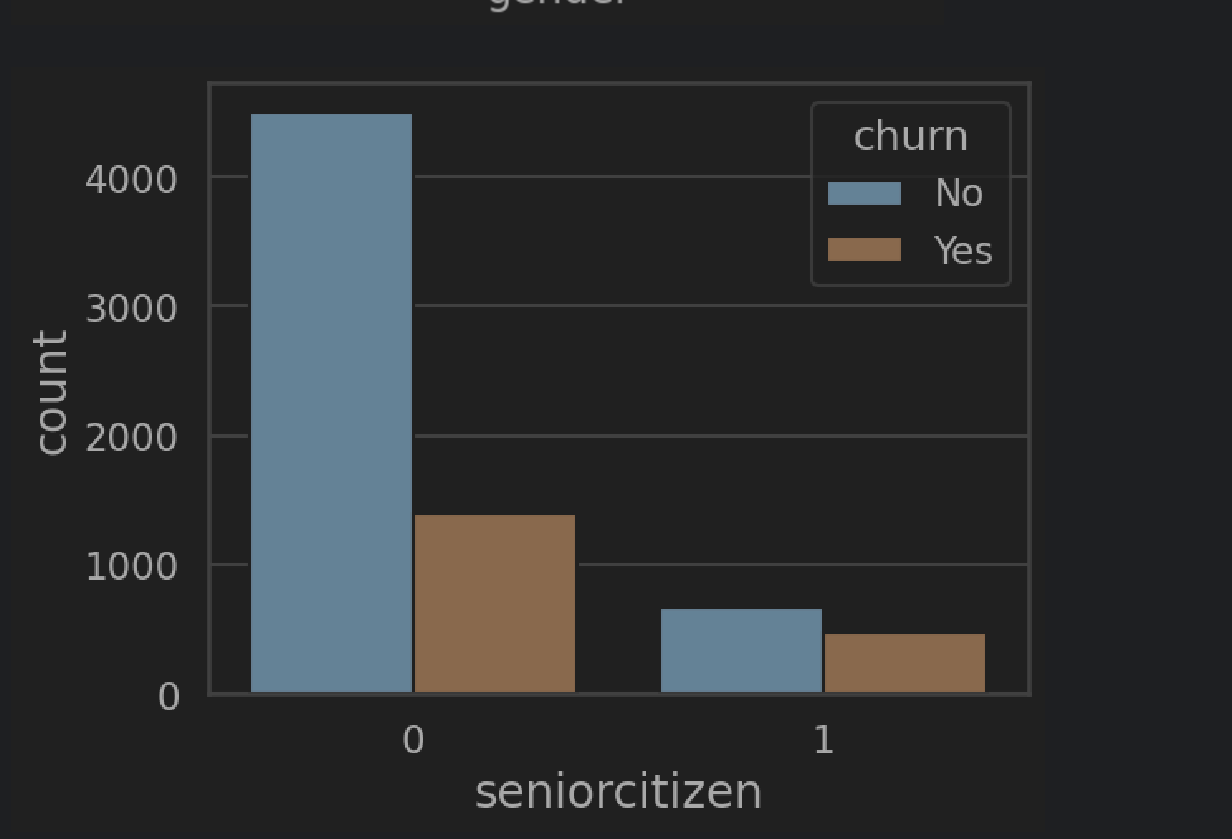


Figure 5.5 Senior Citizen churning rate

The bar graph presented showcases the churn rates among senior citizens, a crucial aspect of customer churn prediction EDA analysis. The graph categorizes customers into senior citizens (1) and non-senior citizens (0), with the count of customers on the y-axis.

A striking feature is the tall blue bar for non-senior citizens, indicating a high number of customers who have not churned, contrasted by a smaller yet significant brown bar representing those who have churned. In comparison, the bars for senior citizens are shorter, denoting fewer customers in this category, but the proportion of churn is visibly higher.

This visual graph suggests that age, specifically being a senior citizen, may be a more influential factor in predicting churn than previously considered. The graph’s clear distinction between the churn rates of senior and non-senior citizens could imply that targeted strategies may be necessary to retain senior customers effectively.

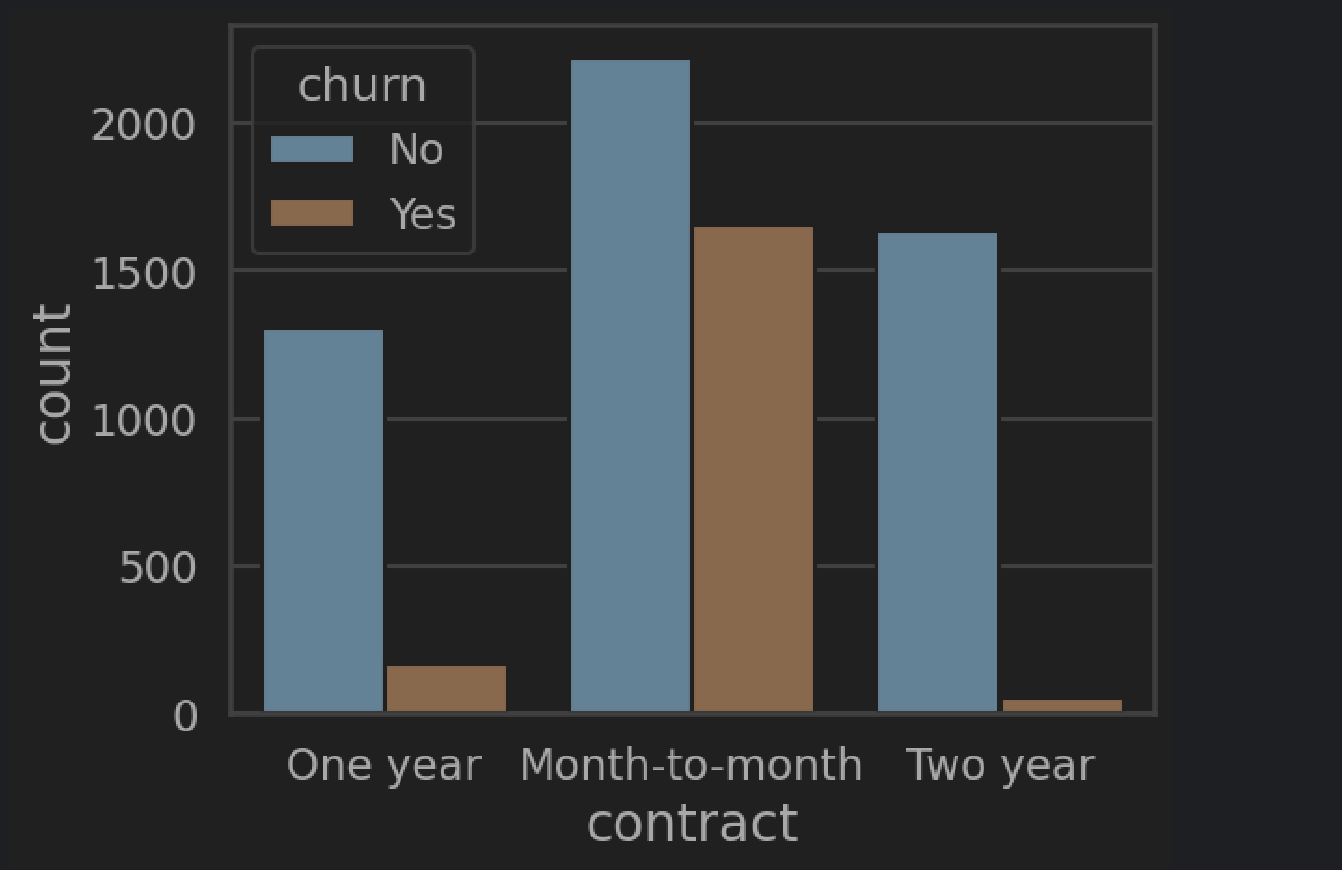


Figure 5.6 Contract type in relation to churning

The bar graph above is an insightful component, focusing on the relationship between contract length and churn rate. It depicts three categories of contract lengths: one year, month-to-month, and two years, with the count of customers who have churned or not churned represented by brown and blue bars, respectively. The graph reveals that customers with a month-to-month contract have the highest churn rate, with over 1500 customers leaving, while those with one-year or two-year contracts show significantly lower churn rates, each with fewer than 500 customers churning. This pattern suggests that contract length is a critical factor in customer retention, with shorter-term contracts likely leading to higher churn.

The above are just some of the insights we looked at in the machine learning data preparation for the model next we needed to prepare data in the proper format for machine learning and that is what we will look at in the next section.

### Preparation for machine model building

In the data preparation phase for machine learning model building, it's crucial to ensure that the data is in a format that the machine learning algorithms can understand and use effectively. This often involves transforming the data in various ways. Looking at our dataset, it's clear that it contains several categorical variables, such as 'gender', 'partner', 'dependents', 'contract', 'paperlessbilling', 'paymentmethod', 'multiplelines', 'internetservice', 'streamingtv', 'streamingmovies', 'onlinesecurity', 'onlinebackup', 'deviceprotection', and 'techsupport'.

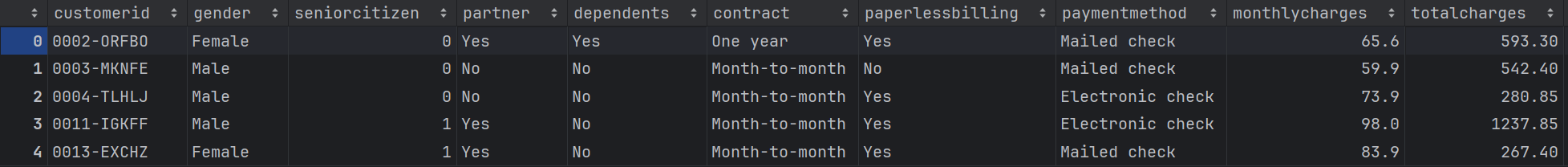


Figure 5.7 Snippet of data structure

These are variables that contain a finite number of discrete categories or classes. Most machine learning algorithms require numerical input and output variables. They cannot work directly with categorical data and require it to be converted into a numerical format. This is where encoding techniques come in.

The technique we used for converting categorical data into a suitable numeric format is using python library called dummies. The get\_dummies function is used for converting the categorical variable into a set of binary variables (also known as dummy variables). For each unique value in the categorical variable, get\_dummies create a new column in the dataset, where the presence of the value is represented by 1/True and absence by 0/False. For instance, in the 'gender' column of your dataset, get\_dummies would create two new columns: 'gender\_Male' and 'gender\_Female'. If the gender of a customer is 'Male', the 'gender\_Male' column had a value of 1 and 'gender\_Female' had a value of 0, and vice versa. And with this we should be able to use our data and find correlation since all values can be read as integers or numeric values

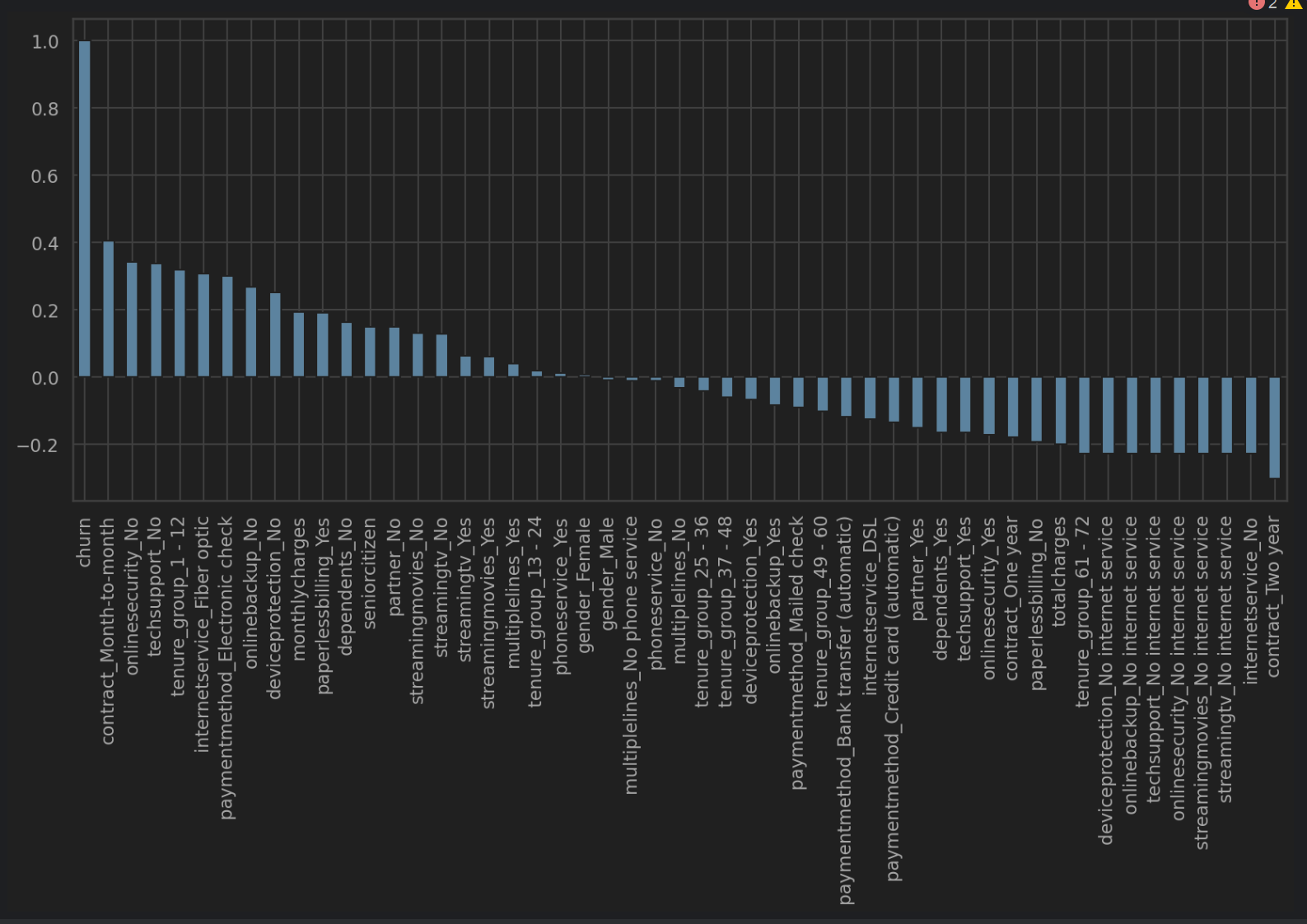


Figure 5.8 Correlation between features

The above graph shows the correlation of our features in relation to churning. At first glance we saw that churn is 1 which indicates it has a correlation of 1 with itself thus explaining the value. As we can see that gender has almost a zero correlation to churning as we saw from the graph earlier. Let’s look at certain features that have strong correlation with churning.

We can saw again as suggested by the graph on contract type it has a strong correlation with churning with month to month having the strongest correlation. This means that we try to predict a user churning based on the contract type of the user. Also, aspects like streaming service and online security shows a higher correlation with our customer churning or not. Now we can see since we can find correlation for each column, we saw that our dummies library was helpful in the readable elements for our machine learning.

### Model building

In this section, we will embark on the process of model building. This is a crucial step in our data analysis where we applied machine learning algorithms to our preprocessed dataset. The goal is to create a predictive model that can accurately determine if a customer churned based on the features we have at our disposal.

We started by splitting our dataset into a training set and a test set. The training set was used to train our model, while the test set was used to evaluate the model's performance. We experimented with various machine learning algorithms, tune their parameters, and compare their performance to select the best model. Throughout this process, we adhered to best practices in machine learning to ensure our model is robust, reliable, and capable of making accurate predictions. This includes techniques such as cross-validation and regularization to prevent overfitting, and metrics such as accuracy, precision, recall, and the F1 score to evaluate model performance.

We started by splitting the data into training and testing sets using train test split library with the test size being 20%. We started by using the DecisionTrees Classifier for our first model. This is because it is easy to interpret and they can mimic the human decision-making process. Also, since our data if full of categorical values it makes even more sense to use DecisionTrees.

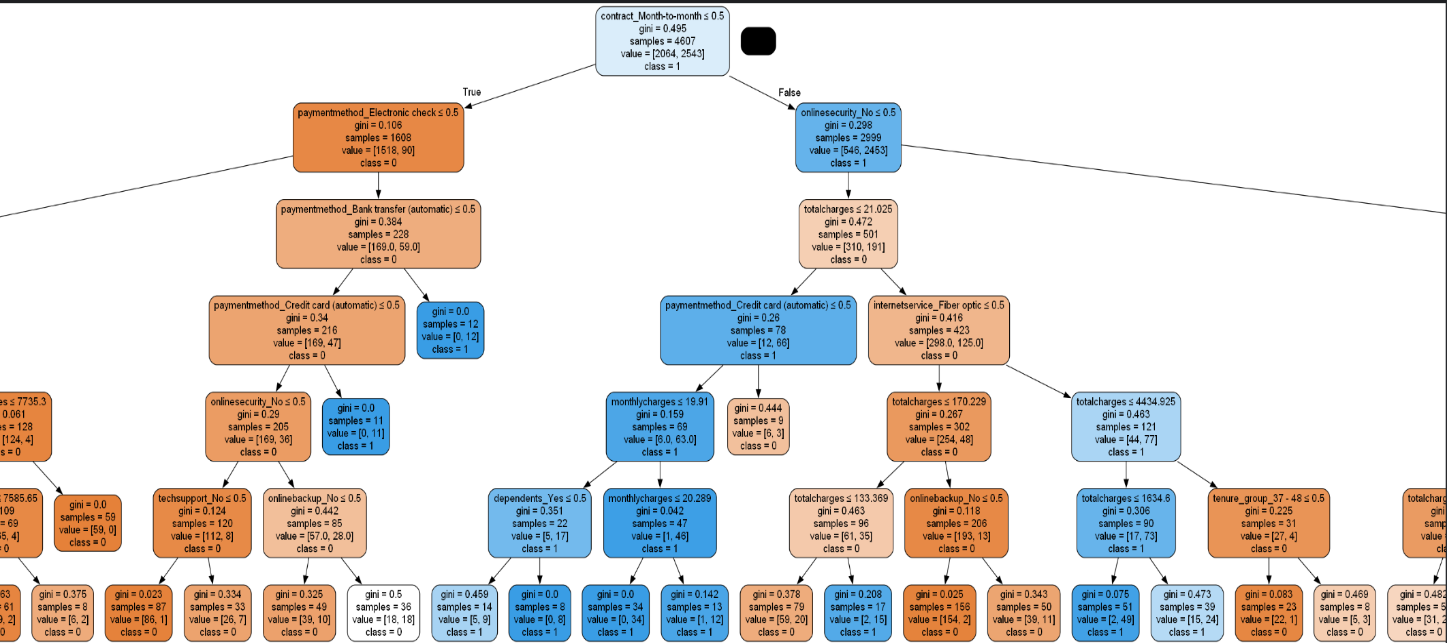


Figure 5.9 Decision Tree Visualization

The above shows the visualization after training our decision tree model and we see how it comes up with the final decision of whether the customer churned or not.

Results for the first model training

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **0** | 0.84 | 0.90 | 0.87 | 1040 |
| **1** | 0.65 | 0.51 | 0.57 | 367 |
| **accuracy** |  |  | 0.80 | 1407 |
| **macro avg** | 0.74 | 0.71 | 0.72 | 1407 |
| **weighted avg** | 0.79 | 0.80 | 0.79 | 1407 |

The classification report provided above is a summary of the precision, recall, and F1-score for a binary classification problem. The precision for class 0 (non-churned customers) is 0.84, which means that when the model predicts a customer did not churn, it is correct 84% of the time. For class 1 (churned customers), the precision is 0.65, meaning the model is correct 65% of the time when it predicts a customer churned.

The recall for class 0 is 0.90, indicating that the model correctly identified 90% of the actual non-churned customers. For class 1, the recall is 0.51, meaning the model correctly identified 51% of the actual churned customers. The F1-score is a weighted average of precision and recall. For class 0, the F1-score is 0.87, and for class 1, it is 0.57. These scores suggest that the model is better at predicting non-churned customers than churned customers. The accuracy of the model is 0.80, meaning the model correctly predicted the churn status for 80% of all customers. The macro average (0.72) and weighted average (0.79) of the F1-score are metrics that summarize the overall performance of the model across all classes. The macro average treats all classes equally, while the weighted average takes into account the imbalance in the class distribution. Both suggest that the model's overall performance is fairly good, but there is room for improvement, especially in correctly identifying churned customers.

As we can remember our data was not balanced and we see that it is good at predicting non-churned customers which means it is the majority of our data. There was a way around these where we used SMOTE Tomek which was used for upscaling our lesser class in order to match the majority class

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| **0** | 0.92 | 0.95 | 0.93 | 544 |
| **1** | 0.95 | 0.93 | 0.94 | 608 |
| **accuracy** |  |  | 0.94 | 1152 |
| **macro avg** | 0.94 | 0.94 | 0.94 | 1152 |
| **weighted avg** | 0.94 | 0.94 | 0.94 | 1152 |

The precision for the '0' class is 0.92, which means that when the model predicts a customer has not churned, it is correct 92% of the time. The recall for the '0' class is 0.95, indicating that the model correctly identified 95% of the actual non-churned customers. The F1-score for the '0' class is 0.93, which is the harmonic mean of precision and recall, providing a single metric that balances both considerations.

For the '1' class, the precision is 0.95, meaning that when the model predicts a customer has churned, it is correct 95% of the time. The recall is 0.93, so the model correctly identified 93% of the actual churned customers. The F1-score is 0.94, again providing a balanced measure of the model's performance for the churned customers. The accuracy of the model is 0.94, meaning that overall, the model made the correct prediction for both churned and non-churned customers 94% of the time. The macro average and weighted average scores are both 0.94, indicating consistent performance across classes when considering different averaging methods. The support values indicate the number of instances of each class in the dataset, with 544 non-churned and 608 churned customers.

We also display a confusion matrix for further insights on the results

|  |  |  |
| --- | --- | --- |
|  | **N** | **P** |
| **T** | 516 | 28 |
| **F** | 45 | 563 |

* True Negative (TN): 516 customers were correctly predicted to not churn. These are the customers who were predicted to stay and did indeed stay.
* False Positive (FP): 28 customers were incorrectly predicted to churn. These are the customers who were predicted to churn but did not.
* False Negative (FN): 45 customers were incorrectly predicted to not churn. These are the customers who were predicted to stay but did churn.
* True Positive (TP): 563 customers were correctly predicted to churn. These are the customers who were predicted to churn and did indeed churn

### Model saving and storage

The model was stored and saved using the pickle module in Python. pickle is used for serializing (also called pickling) and de-serializing Python object structures. Serialization refers to the process of converting an object's state to a byte stream, and the opposite operation, creating object from a byte stream, is known as deserialization. In our case, the trained model model\_rf\_smote is saved to a file named "model.sav". This is done using pickle.dump(), which serializes the object and writes it to the specified file. This process allows the model to be stored so that it can be used later without needing to be retrained. Later, the model is loaded from the file using pickle.load(). This function deserializes the file content and recreates the original object, in this case the trained model. The loaded model is then used to score the test data, demonstrating that the model was successfully stored and loaded.

## Testing results and verdict

1. Test Case

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Test ID** | **Related Requirement** | **Check** | **Pre-Condition** | **Test data** | **Priority level** |
| **1** | FRQ1 | Does the data pipeline extract data from the csv files | CSV files should be available in the directory | Customer\_churn.csv | High |
| **2** | FRQ2 | Does the data pipeline transform data from csv files | Data should have been extracted from csv file | Customer churn data frame | High |
| **3** | FRQ3 | Does data pipeline load data into the database | Data should have been transformed | Customer churn data frame | High |
| **4** | FRQ4 | Does machine learning model able to query from the data base | Data should have been loaded to data base | Customer churn SQL query | High |
| **5** | FRQ5 | Does the Airflow pipeline execute without errors | Airflow should be properly installed and configured | Airflow DAG configuration | High |
| **6** | FRQ6 | Does the Flask application start successfully | Flask should be properly installed and configured | Flask application configuration | High |
| **7** | FRQ7 | Does the machine learning model predict churn accurately | Model should have been trained successfully | Test data set | High |
| **8** | FRQ8 | Does the Flask application return predictions successfully | Model should have been loaded into the Flask application | Customer details input data | High |
| **9** | FRQ9 | Does the Flask application handle invalid queries gracefully | Flask application should be running | Invalid customer details input data | Medium |
| **10** | FRQ10 | Does the data pipeline handle missing or incorrect CSV files gracefully | CSV files should be available in the directory | Missing or incorrect CSV files | Medium |
| **11** | FRQ11 | Does the machine learning model handle missing or incorrect data gracefully | Data should have been loaded into the model | Incorrect or missing data set | Medium |
| **12** | FRQ12 | Does the Airflow pipeline handle task failures gracefully | Airflow should be properly installed and configured | configured  Airflow DAG configuration with task designed to fail | Low |
| **13** | FRQ13 | Does the machine learning model handle imbalanced data sets gracefully | Data should have been loaded into the model | Imbalanced data set | Low |
| **14** | FRQ14 | Does the data pipeline handle large CSV files without running out of memory | Large CSV files should be available in the directory | Large CSV files | Low |

1. Test results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test ID** | **Expected results** | **Actual Results** | **Status** | **Remarks** |
| **1** | Data is successfully extracted from the csv file | Data was extracted and stored in data frame object from the csv | Success | Good performance |
| **2** | Data is successfully transformed in the pipeline | Data was transformed corrected to correct format needed for database and cleaned | Success | Good performance |
| **3** | Data is successfully loaded into the database | Data was loaded to the database and created database structured described in the pipeline | Success | Good performance |
| **4** | Data scientist is able to connect to the database and query from notebook | Jupyter notebook is able to connect to database and query it | Success | Good performance |
| **5** | Airflow pipeline executes without errors | Airflow executed scheduled dags with no failures and reports on success | Success | Good performance |
| **6** | Flask application starts successfully |  | Success | Good performance |
| **7** | Machine learning model predicts churn accurately | Machine learning model predicts with accuracy of 95% with f1-score of 94% | Success | Good performance |
| **8** | Flask application returns predictions successfully | Chance of churning is returned by the flask application | Success | Good performance |
| **9** | Flask application handles invalid queries gracefully | Flask can only trigger warnings | Success | Good performance |
| **10** | Data pipeline handles missing or incorrect CSV files gracefully | Airflow reports on any errors and attempts to run again | Success | Good performance |
| **11** | Machine learning model handles missing or incorrect data gracefully | Machine learning model has no missing values since data pipeline handled all cleaning | Success | Good performance |
| **12** | Airflow pipeline handles task failures gracefully | Airflow retries tasks and sends error logs or notification for catastrophic failures | Success | Good performance |
| **13** | Machine learning model handles imbalanced data sets gracefully | Machine learning uses SMOTEEK to handle imbalanced data set | Success | Good performance |
| **14** | Data pipeline handles large CSV files without running out of memory | Model not made for high data volumes and takes too much resources | Success | Good performance |

The tests had been successful, leading to the establishment of a robust data pipeline. This pipeline was efficient in the Extract, Transform, Load (ETL) process, and the scheduled runs were executed without any hitches. The machine learning model, although it did not reach the targeted level of accuracy, still performed fairly well. It achieved an accuracy of 95%, which, while not the ultimate goal, was still a high mark. This level of accuracy allowed the model to predict a majority of the data correctly. In essence, despite not reaching the desired accuracy, the model and the data pipeline proved to be effective and reliable in handling and predicting the data.

# CONCLUSIONS, RECOMMENTATIONS AND FUTURE WORKS

## Conclusions

In the modern era of information technology, data has emerged as a crucial asset, akin to the importance of oil in the industrial age. The effective utilization of data has become a necessity for organizations to maintain competitiveness and generate value. However, the raw nature of data necessitates transformation into meaningful insights, a task typically undertaken by professionals such as Data Analysts, BI Analysts, and Data Scientists.

This highlights the critical need for robust data integration within businesses to extract actionable insights. One of the significant challenges businesses face today is customer churn, where consumers disengage from online products or services. This phenomenon leads to reduced revenue and sluggish growth. While the reasons behind customer churn may not be straightforward, data can provide crucial insights into understanding and predicting this occurrence. This underscores the importance of proper data integration in business operations, using data not only to comprehend the issue but also to anticipate and mitigate customer churn.

## Recommendations

To effectively address customer churn, businesses should invest in creating a robust data ecosystem that supports a predictive system for customer churn. This involves integrating various data sources, cleaning and transforming raw data, and applying advanced analytics and machine learning techniques to predict churn. Businesses should also invest in training their staff, particularly Data Analysts, BI Analysts, and Data Scientists, to effectively use these systems and interpret the results. Also, they need to know the importance of data engineer this is to know where the business stands in the case of data needs and data maturity stage.

## Future works

For future works and improvements, we can scale the system to deal with big data and acquire data from sensors, this will require the use of NoSQL for better performance and Apache spark for handling the large amounts of data. Also building a machine learning pipeline in cases where a lot of data preparation and modelling needs to be done

# References

Argelaguet, R. C. (2021). *Computational principles and challenges in single-cell data integration.* Nature biotechnology.

Bhatia, P. (2019). Data mining and data warehousing. *Principles and practical techniques*.

Çelik, O. &. (2019). Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments*, 30-38.

de Lima Lemos, R. A. (2022). Propension to customer churn in a financial institution: A machine learning approach. *Neural Computing and Applications*, 11751-11768.

Fedushko, S. &. (2022). E-commerce customers behavior research using cohort analysis: A case study of COVID-19. *Journal of Open Innovation: Technology, Market, and Complexity*, 12.

Fonseca, G. &. (2023). Overcoming the challenges of data integration in ecosystem studies with machine learning workflows: an example from the Santos project. *Ocean and Coastal Research*, 71.

Hira, S. &. (2023). Automated heuristic based context dependent ETL process to generate multi‐dimensional model for tabular data. *Concurrency and Computation: Practice and Experience*, 35.

J. M., M. J. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences*, 118.

Joe Reis, M. H. (2022). *Fundamentals of Data Engineering.* Sebastopol, CA: O'Reilly Media, Inc.

Kosvyra, A. F. (2021). Towards data integration for AI in cancer research. *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2054-2057.

McGuirk, M. (2023). Performing web analytics with Google Analytics 4: a platform review. *Journal of Marketing Analytics*, 1-15.

Momin, S. B. (2020). Prediction of customer churn using machine learning. *EAI International Conference on Big Data Innovation for Sustainable Cognitive Computing: BDCC 2018*, 203-212.

Munappy, A. R. (2020). From ad-hoc data analytics to dataops. *In Proceedings of the International Conference on Software and System Processes*, 165-174.

Murphy, G. (2022). The 5 Tenets of Data Integration. *The 5 Tenets of Data Integration*, 1-3.

Nonis, P. R. (2022). Customer retention and addressing customer churn through Predictive Analytics in telecom industry. *Doctoral dissertation*.

Oluwatoyin, A. M. (2022). Customer Churn Prediction in Banking Industry Using Power Bi. *In Proceedings of Third International Conference on Computing, Communications, and Cyber-Security:*, 767-774.

Raj, A. B. (2020). Modelling data pipelines. *2020 46th Euromicro conference on software engineering and advanced applications*, 13-20.

Ralph Kimball, M. R. (2019). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.* John Wiley and Sons, Inc.

Yu, M. L. (2020). A new framework for online testing of heterogeneous treatment effect. *Proceedings of the AAAI Conference on Artificial Intelligence*, 10310-10317.

.

Appendix

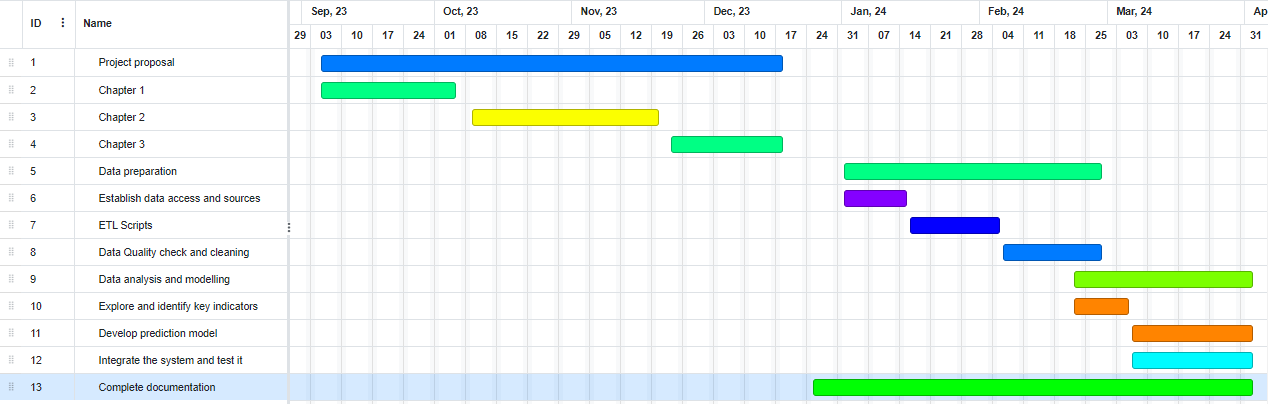


Figure 0.1 Gant Chart