**Predictive Analysis of Newspaper Subscribers Using Big Data**

**TEAM #3**

|  |  |
| --- | --- |
| Aakash Boenal | G01413569 |
| Kiranmai Reddy Varakantam | G01282111 |
| Meghana Katta | G01408184 |
| Rishitha Madipelli | G01410830 |
| Shreya Gowrishetty | G01406548 |

Huajun Zhang

George Mason University

AIT-614: Big Data Essentials

October 24th, 2023

**TABLE OF CONTENTS**

CONTENTS PG NO.

[INTRODUCTION 3](#_Toc149083915)

[RELATED WORK 3](#_Toc149083916)

[OBJECTIVES 3](#_Toc149083917)

[PROPOSED SELECTED DATASET 4](#_Toc149083918)

[DESCRIPTION OF PROPOSED SYSTEM 4](#_Toc149083919)

[PROPOSED DEVELOPMENT PLATFORMS 5](#_Toc149083920)

[PROJECT TASKS AND TIMELINE 5](#_Toc149083921)

[REFERENCES 6](#_Toc149083923)

[APPENDIX 6](#_Toc149083924)

# **INTRODUCTION**

In today's digital age, newspapers are adapting to changing readership patterns, and understanding subscriber behavior is essential for retaining and acquiring readers. Big data analytics can provide valuable insights into subscriber preferences, enabling newspapers to make data-driven decisions for improved customer engagement and content delivery. This project's scope includes the utilization of a Newspaper Subscribers dataset for predictive modeling, leveraging big data technologies to analyze the data, and contributing to the field of media analytics.

Newspaper companies worldwide are experiencing a shift from traditional print to digital subscriptions. To adapt to this changing landscape, understanding factors that influence subscriber retention and churn becomes crucial. Our project's primary objective is to build a predictive model that can provide insights into these factors and aid in making data-driven decisions.

# **RELATED WORK**

In the field of predictive analytics, previous research has explored predictive modeling in various industries, including media and publishing. Studies have highlighted the importance of understanding customer demographics, behavior, and preferences for reducing churn rates. However, this project's approach differs in terms of the specific dataset used, the methods employed, and the focus on applying PySpark and Spark MLlib for scalable analysis. By leveraging distributed computing and big data technologies, this project aims to offer more robust and efficient predictions for newspaper subscriber behavior.

# **OBJECTIVES**

The main objectives of the project are as follows:

1. **Data Preprocessing and Cleaning**: We aim to prepare the Newspaper Subscribers dataset for analysis by ensuring it is free from errors and inconsistencies. This is crucial for the accuracy of our predictive model.
2. **Exploratory Data Analysis (EDA)**: We will perform EDA to uncover valuable insights into subscriber behavior. By visualizing and summarizing the data, we can identify patterns, trends, and potential areas of interest.
3. **Develop a Predictive Model**: Our primary goal is to build a predictive model that can determine subscriber preferences and forecast subscriber churn. This model will be a cornerstone of our project.
4. **Implementation with PySpark and Spark MLlib**: We will leverage PySpark and Spark MLlib to implement the predictive model, ensuring it can handle large-scale data for scalability and efficiency.
5. **Model Evaluation and Fine-Tuning**: We will rigorously assess the performance of our predictive model and fine-tune it to enhance accuracy. This iterative process will help us create a highly reliable model.
6. **Recommendations for Content Delivery and Retention**: Based on the insights gained from our analysis and predictive model, we will provide recommendations for targeted content delivery and strategies to improve subscriber retention. This ensures our project delivers practical value to stakeholders.

# **PROPOSED SELECTED DATASET**

We have selected a dataset that contains over 15,000 records of newspaper subscribers. The dataset is rich in demographic information, including household income, home ownership, age range, and more. Our dataset meets the project's criteria with more than 10 original features and over 10,000 data records. These could include things like age, where people live, how long they've been subscribing, how often they read the newspaper, and more. All of this information will help us create computer models that can predict things about subscribers, like whether they might stop subscribing.

In summary, the dataset provides the necessary raw materials to construct and train predictive models that can unearth insights into subscriber behavior, preferences, and potential churn indicators. It is essential to emphasize that the dataset's quality and relevance are crucial in achieving the project's objectives, and thorough data preprocessing and cleaning will be conducted to ensure its suitability for analysis.

# **DESCRIPTION OF PROPOSED SYSTEM**

A diagram of a model

Description automatically generated

Fig 1: Proposed System Architecture

# **PROPOSED DEVELOPMENT PLATFORMS**

The proposed development platforms for our project will encompass software resources to ensure a robust and scalable environment. We will utilize the following components:

1. **NoSQL Database:** We will employ a NoSQL database, such as Databricks DBFS, for storing and managing our dataset. NoSQL databases are well-suited for handling large volumes of unstructured or semi-structured data, making them a vital component of our data analysis pipeline.
2. **Big Data Engines:**
   * **PySpark**: PySpark, a Python library for Apache Spark, will serve as one of our primary big data engines. It is ideal for distributed data processing and machine learning tasks, allowing us to efficiently handle big data analytics.
   * **Spark MLlib**: We will also leverage Spark MLlib, a machine learning library within Apache Spark, to develop and deploy machine learning models. It provides a range of algorithms and tools for predictive analytics.
3. **Big Data Platform:** We will explore cloud-based Big Data platforms, such as AWS (Amazon Web Services), Azure, Apache Hadoop, or Google Cloud Platform. These platforms offer scalable and cost-effective infrastructure, making it easier to manage and analyze large datasets. The choice of platform will depend on the team's familiarity and available resources.

# **PROJECT TASKS AND TIMELINE**

# Our project tasks and timeline (as of now) are outlined as follows:

|  |  |  |
| --- | --- | --- |
| **Task Description** | **Team Members** | **Timeline** |
| Data Preprocessing and Cleaning | Aakash | 2 weeks |
| Exploratory Data Analysis (EDA) | Kiranmai | 2 weeks |
| Predictive Model Development | Meghana | 2 weeks |
| PySpark and Spark MLlib Implementation | Rishitha | 2.5 weeks |
| Model Evaluation and Fine-Tuning | Shreya | 2.5 weeks |

# **REFERENCES**

[1] geiler, l. (10, May 2022). *NewspaperChurn .* Retrieved from Open ML: <https://api.openml.org/d/44226>

[2] Belchior, L. M. (2022, November). *Churn Prediction in Online Newspaper Subscriptions.* Retrieved from Run Repositorio Universidade Nova: <https://run.unl.pt/bitstream/10362/148914/4/TGI1079.pdf>

[3] Peña, V. C., Malthouse, E. C., & Mersey, R. D. (2023). Churning off the news: An analysis of newspaper subscriber churn across digital devices. Newspaper Research Journal, 44(2), 190-205. <https://doi.org/10.1177/07395329231167374>

# **APPENDIX**

Below are the first few rows of the dataset:

A screenshot of a computer

Description automatically generated