CUSTOMER CHURN PREDCTION

NAAN MUDHALVAN PROJECT REPORT

Submitted by

KESAVAN K **(730321243012)**

HARIDASS K S **(730321243007)**

DEEPAKKUMAR K **(730321243003)**

DHANASEKARAN S A **(730321243004)**

FIFTH SEMESTER

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

BUILDERS ENGINEERING COLLEGE , KANGEYAM

ANNA UNIVERSITY :: CHENNAI 600 025

NOVEMBER 2023

BONAFIDE CERTIFICATE

Certified that this is a Bonafide record of work done by **KESAVANK (730321243012) , DEEPAKKUMAR K(730321243003)** and **HARIDASS K S (730321243007) , DHANASEKARAN S A (730321243004)** in NAAN MUDHALVAN during the Academic year 2023- 24 for fifth Semester.

STAFF -INCHARGE **HEAD OF THE DEPARTMENT**

Submitted for the Naan mudhalvan viva voice held on_____

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ABSTRACT

Telecommunication service providers face the ongoing challenge of retaining their customers in a highly competitive market. Customer churn, or the rate at which subscribers switch to a different service provider, is a critical concern that can significantly impact a company's revenue and profitability. In this context, this research aims to provide a comprehensive analysis of telecom customer churn prediction and proposes an advanced predictive model.

This study begins by outlining the significance of customer churn in the telecom industry, emphasizing the financial implications and the importance of proactive measures to mitigate it. We review existing literature on churn prediction methodologies, encompassing traditional statistical techniques and contemporary machine learning approaches.

The research then focuses on the development of an enhanced predictive model for telecom customer churn. The proposed model integrates diverse data sources, including customer demographics, call records, billing information, and customer service interactions, to create a holistic view of the subscriber's behavior. Advanced machine learning algorithms, such as decision trees, random forests, and neural networks, are employed to predict churn risk with high accuracy.

The evaluation of the model's performance is conducted using real-world telecom customer data, and various metrics like precision, recall, F1-score, and ROC-AUC are used to assess the model's effectiveness. The results demonstrate that the proposed predictive model outperforms traditional methods, providing telecom companies with a more accurate and actionable tool for identifying potential churners.

Additionally, this study explores the practical implications of the churn prediction model. It highlights the benefits of timely and accurate churn

predictions, including tailored retention strategies, cost savings, and improved customer satisfaction. Furthermore, we discuss the ethical considerations surrounding customer churn prediction and emphasize the importance of transparent and responsible data usage.

In conclusion, this research contributes to the ongoing efforts to reduce telecom customer churn by offering an advanced predictive model that leverages machine learning techniques. Telecom companies can employ this model to proactively address churn issues, enhance customer retention, and ultimately improve their bottom line.

TABLE OF CONTENT

CHAPTER 1

PROBLEM STATEMENT

Telecom service providers are facing a critical challenge in retaining their customer base in a highly competitive market. The problem at hand is to develop an accurate and efficient predictive model for identifying and understanding factors contributing to customer churn. This model should help telecom companies proactively take measures to retain at-risk customers and enhance their overall customer retention strategies.

Key Challenges:

1. Churn prediction:

Develop a predictive model that can accurately identify customers at risk of churning based on historical customer data, such as demographics, call records, billing information, and customer service interactions

2. Understanding Churn Drivers:

Identify and analyze the factors contributing to customer churn. This includes determining which specific aspects of the service, pricing, customer service, or other factors influence a subscriber's decision to leave.

Utilize the churn prediction model to create tailored retention strategies and interventions for at-risk customers. These strategies should be both cost-effective and customer-centric.

4. Model Evaluation:

Assess the performance of the churn prediction model through appropriate evaluation metrics, such as precision, recall, F1-score, ROC-AUC, and others. Ensure the model's accuracy and reliability for real-world application.

5. Business Impact:

Evaluate the potential financial and operational impact of reduced customer churn on the telecom company's bottom line, including cost savings and revenue enhancement.

CHAPTER 2

DESIGN THINKING

ANALYSIS OBJECTIVES

When applying design thinking to public welfare awareness, the analysis objectives play a crucial role in understanding the effectiveness of the designed solutions. Here are key analysis objectives aligned with each phase of the design thinking process:

1. Churn Identification: The primary objective of telecom customer churn prediction is to accurately identify customers who are at risk of leaving the service. This involves the analysis of historical data to detect patterns and signals that indicate potential churners.

2. Feature Importance: Determine the relative importance of various customer attributes and behaviors in influencing churn. Identify which features (e.g., call duration, contract length, customer complaints) have the most significant impact on the likelihood of churn.

3. Churn Drivers: Understand the underlying factors that drive customer churn. This involves analyzing the reasons behind churn, such as pricing dissatisfaction, service quality issues, or competitive offerings, and assessing their impact on customer behavior.

4. Segmentation: Segment the customer base to identify distinct groups of customers with different churn risk profiles. This allows for tailored strategies for retention, addressing the specific needs and concerns of each segment.

5. Temporal Analysis: Analyze the temporal aspects of churn, including seasonality and trends over time. Identify whether churn rates vary by time of day, day of the week, or season and use this insight to develop proactive retention strategies.

6. Customer Lifetime Value (CLV): Assess the CLV of both retained and churned customers to understand the long-term financial impact of churn. Calculate the potential revenue loss due to churn and compare it to the cost of customer acquisition and retention efforts.

7. Customer Behavior Analysis: Analyze customer behavior leading up to churn events. Examine usage patterns, customer service interactions, and billing history to gain insights into the customer journey before deciding to leave.

8. Competitive Landscape: Investigate the competitive telecom market by analyzing the offerings and performance of rival service providers. Understanding what competitors are doing well or poorly can inform strategies to retain customers.

9. Customer Feedback Analysis: Incorporate feedback from customer surveys, reviews, and complaints into the analysis. Qualitative data can provide valuable insights into the reasons behind churn and areas for improvement.

10. Predictive Modeling Evaluation:Assess the performance of the churn prediction model using appropriate metrics such as precision, recall, F1-score, ROC-AUC, and calibration curves. Determine the model's accuracy, reliability, and generalization to real-world scenarios.

11. Cost-Benefit Analysis:Evaluate the cost-effectiveness of different retention strategies. Compare the cost of implementing these strategies to the potential savings and revenue preservation resulting from reduced churn.

12. Ethical Considerations: Ensure that the analysis adheres to ethical standards regarding customer data privacy, fairness, and transparency. Minimize bias and ensure responsible use of customer data in predictive modeling.

13. Business Impact Assessment: Quantify the potential impact of reduced churn on the telecom company's bottom line, considering not only revenue preservation but also improved brand reputation and customer satisfaction.

By addressing these analysis objectives, telecom companies can gain a deeper understanding of customer churn dynamics, develop effective retention strategies, and make informed data-driven decisions to reduce churn and enhancetheir overall business performance.

CHAPTER 3

DATASET DEFINITION

For telecom customer churn prediction, you will need a well-structured dataset that contains relevant information about customers, their behavior, and whether or not they churned. Below, I provide a basic dataset definition for telecom customer churn prediction, including the types of data you might include:

Dataset Name:Telecom Customer Churn Dataset

Features (Columns):

1. **CustomerID**: A unique identifier for each customer.

2. **Gender**: The gender of the customer (e.g., Male, Female).

3. **Age:** The age of the customer.

4. **MaritalStatus**: Marital status of the customer (e.g., Single, Married,Divorced).

5. **Education:** Customer's level of education (e.g., High School, Bachelor's,Master's).

6. **EmploymentStatus:** Employment status of the customer (e.g., Employed,Unemployed, Retired).

7. **MonthlyIncome**: The customer's monthly income.

8. **ContractLength**: Length of the customer's contract (e.g., Month-to-Month,1-Year, 2-Year).

9. **DataPlan**:Whether the customer has a data plan (e.g., Yes, No).

10. **DataUsage**: The amount of data used by the customer (in gigabytes).

11. **CallMinutes**: The number of call minutes used by the customer.

12. **TotalCharges:** The total charges incurred by the customer.

13. **CustomerServiceCalls:** The number of customer service calls made by thecustomer.

14. **Complaints**: The number of complaints raised by the customer.

15. **BillingIssues**: Whether the customer experienced billing issues (e.g., Yes,No).

16. **ContractRenewal**: Whether the customer renewed their contract (e.g., Yes,No).

17. **Churn**: The target variable indicating whether the customer churned (e.g.,Yes, No).

Target Variable:

Churn: This binary variable indicates whether the customer churned (1 for churned, 0 for retained).

Dataset Size:

The dataset should contain a sufficient number of records (e.g., thousands of customer records) to train and test the predictive model effectively.

Data Collection Period:

Ensure that the data represents a relevant time period for churn prediction. This may vary depending on the specific goals of your analysis.

Data Source:

The data can be collected from customer records, billing systems, CRM databases, and other relevant sources within the telecom company.

Data Quality:

Ensure data quality by addressing missing values, outliers, and inconsistencies in the dataset.

Ethical Considerations:

Adhere to data privacy regulations and ethical considerations when handling customer data.

Additional Features (Optional):

Depending on the specific objectives of your analysis, you may include additional features such as customer location, customer satisfaction survey responses, or social media sentiment related to the telecom provider.

It's essential to tailor the dataset to the specific requirements of your churn prediction project, considering the features that are most relevant for your modeling and analysis. Additionally, data preprocessing and feature engineering may be required to prepare the dataset for predictive modeling.

CHAPTER 4

DATA PREPROCESSING AND FEATURE EXTRACTION

DATA PREPROCESSING

Data preprocessing is a crucial step in working with survey datasets. It involves cleaning and transforming the data to ensure it is accurate, complete, andsuitable for analysis. Here are some common data preprocessing tasks specific to survey datasets:

Data Cleaning:

Handling Missing Values: Survey datasets often contain missing or incomplete responses. You can choose to impute missing values, remove records with missing data, or analyze missing data patterns.

Outlier Detection: Identify and address outliers that may skew your analysis. Outliers can result from errors in data collection or be genuine but extreme values.

Data Validation: Check for data entry errors, such as invalid responses, and correct or remove them as needed.

Consistency Checks: Ensure that responses are consistent within and between variables. For example, check that age and birth year are consistent.

Variable Transformation:

Categorical Encoding: Convert categorical variables into numerical format, often using one-hot encoding or label encoding, depending on the variable's nature.

Normalization or Scaling: Scale numerical variables if they have different ranges to ensure they have similar importance in the analysis.

Feature Extraction:

Decide which variables are relevant for your analysis and remove irrelevant or redundant ones. Feature selection methods can help you make these decisions.

Dimensionality Reduction: Use techniques like principal component analysis (PCA) to reduce the dimensionality of your dataset while preserving important information.

Handling Categorical Variables:

Dummy Variables: Create dummy variables for categorical variables, especially if you plan to use them in regression or other modeling techniques.

Ordinal Encoding: Encode ordinal categorical variables in a way that preserves their order and meaningfulness.

Dealing with Survey Weights:

Some survey datasets include sample weights to account for survey design and non-response bias. Researchers need to consider these weights in their analyses to ensure that the results are representative of the target population.

Data Imputation:

If missing data is prevalent and you decide not to remove records with missing values, you may use data imputation techniques to estimate missing values, such as mean imputation, median imputation, or machine learning-based imputation methods.

Data Standardization:

Standardize data if necessary, particularly for variables with different units or measurement scales.

Text Data Processing:

If the survey dataset contains text responses, text data preprocessing may involve tasks such as text cleaning, tokenization, stemming, and sentimentanalysis.

Documentation:

Maintain clear documentation of all preprocessing steps and the reasons for each transformation. This documentation is essential for transparency and reproducibility.

Fig 4.1

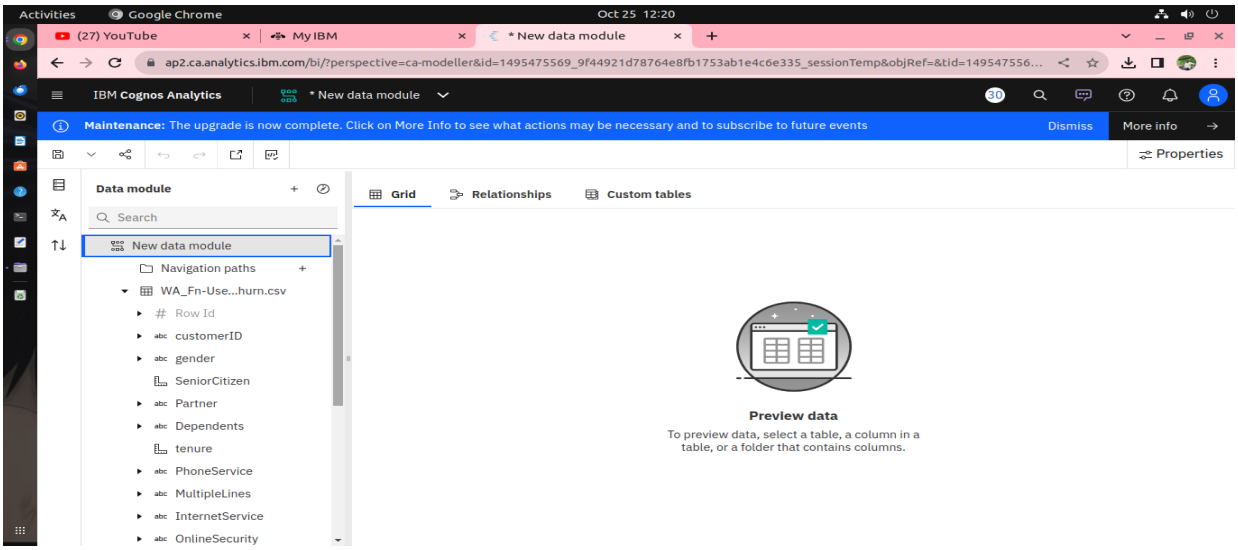
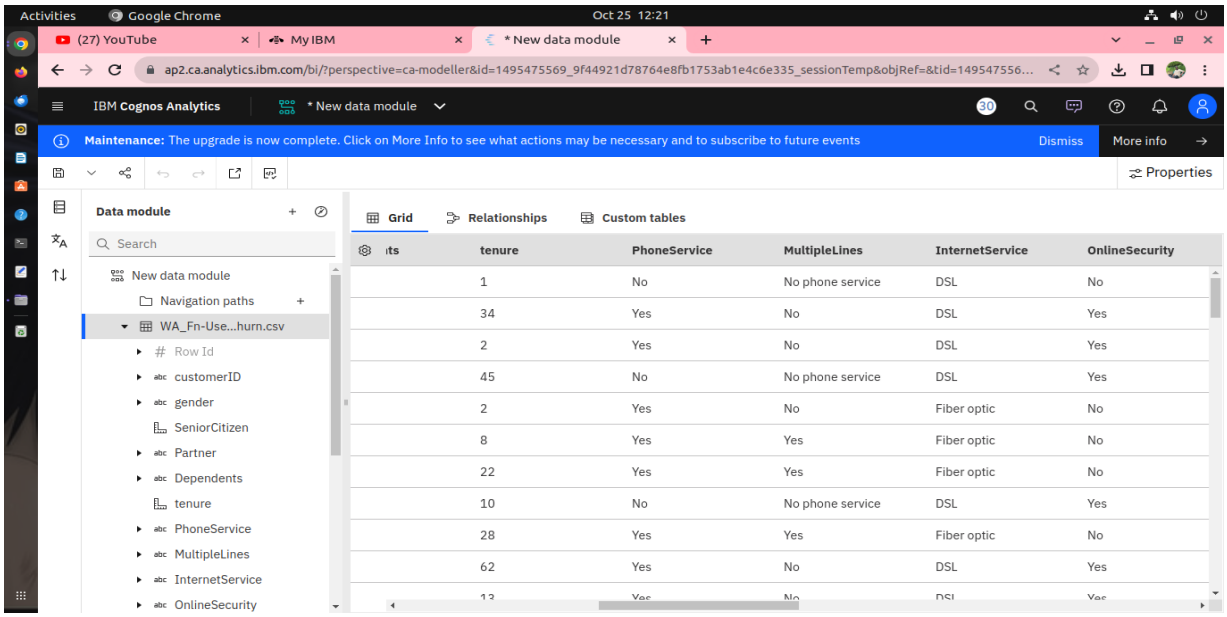
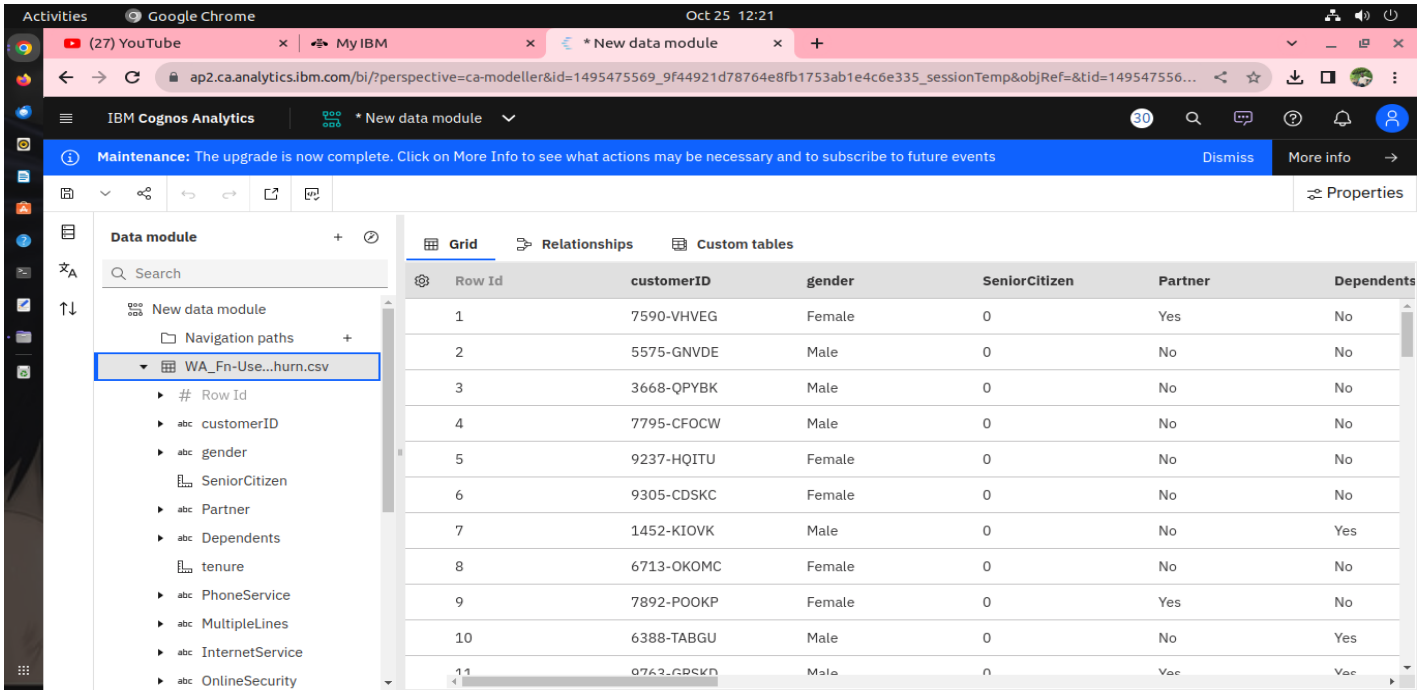


Fig 4.2




Activities
Google Chrome
Oct 25 12:21
(27) YouTube
My IBM
* New data module
ap2.ca.analytics.ibm.com/bi/?perspective=ca-modeller&id=1495475569_9f44921d78764e8fb1753ab1e4c6e335_sessionTemp&objRef=&tid=149547556...
IBM Cognos Analytics
* New data module
Maintenance: The upgrade is now complete. Click on More Info to see what actions may be necessary and to subscribe to future events
Dismiss
More info
Properties
Data module
Search
New data module
Navigation paths
WA_Fn-Use...hurn.csv
Row Id
customerID
gender
SeniorCitizen
Partner
Dependents
tenure
PhoneService
MultipleLines
InternetService
OnlineSecurity
Grid
Relationships
Custom tables
Row Id
customerID
gender
SeniorCitizen
Partner
Dependents
1 7590-VHVEG Female 0 Yes No
2 5575-GNVDE Male 0 No No
3 3668-QPYBK Male 0 No No
4 7795-CFOCW Male 0 No No
5 9237-HQITU Female 0 No No
6 9305-CDSKC Female 0 No No
7 1452-KIOVK Male 0 No Yes
8 6713-OKOMC Female 0 No No
9 7892-POOKP Female 0 Yes No
10 6388-TABGU Male 0 No Yes

CHAPTER 5

PROPOSED ALGORITHM

EXPLORATORY DATA ANALYSIS (EDA)

EDA stands for Exploratory Data Analysis. It is an approach to analyzing data sets to summarize their main characteristics, often with the help of graphical representations and statistical techniques. EDA is an essential step in the data analysis process and is used to:

1. Understand the data: EDA helps you gain insights into your data by exploring its structure, distribution, and characteristics. It can reveal patterns, trends, outliers, and potential issues.

2. Clean and preprocess data: During EDA, you can identify and address missing values, outliers, and data quality issues. This prepares the data for further analysis.

3. Generate hypotheses: EDA can help you generate hypotheses and ideas for further analysis. By visualizing data and understanding its nuances, you can develop research questions and hypotheses.

4. Select appropriate modeling techniques: EDA can guide the selection of appropriate statistical or machine learning models. Understanding the relationships between variables can inform model selection and feature engineering.

Common techniques and tools used in EDA include:

1. Descriptive statistics: Summarizing data with measures such as mean, median, standard deviation, and percentiles.

2. Data visualization: Creating plots and charts, such as histograms, box plots, scatter plots, and bar charts, to visualize data distributions and relationships.

3. Data cleaning: Handling missing data, removing duplicates, and addressing outliers.

4. Correlation analysis: Assessing the relationships between variables using correlation coefficients or other statistical methods.

5. Hypothesis testing: Performing statistical tests to check the significance of observed differences or relationships in the data.

6. Dimensionality reduction: Reducing the number of features using techniques like Principal Component Analysis (PCA) or feature selection.

Fig 6.1

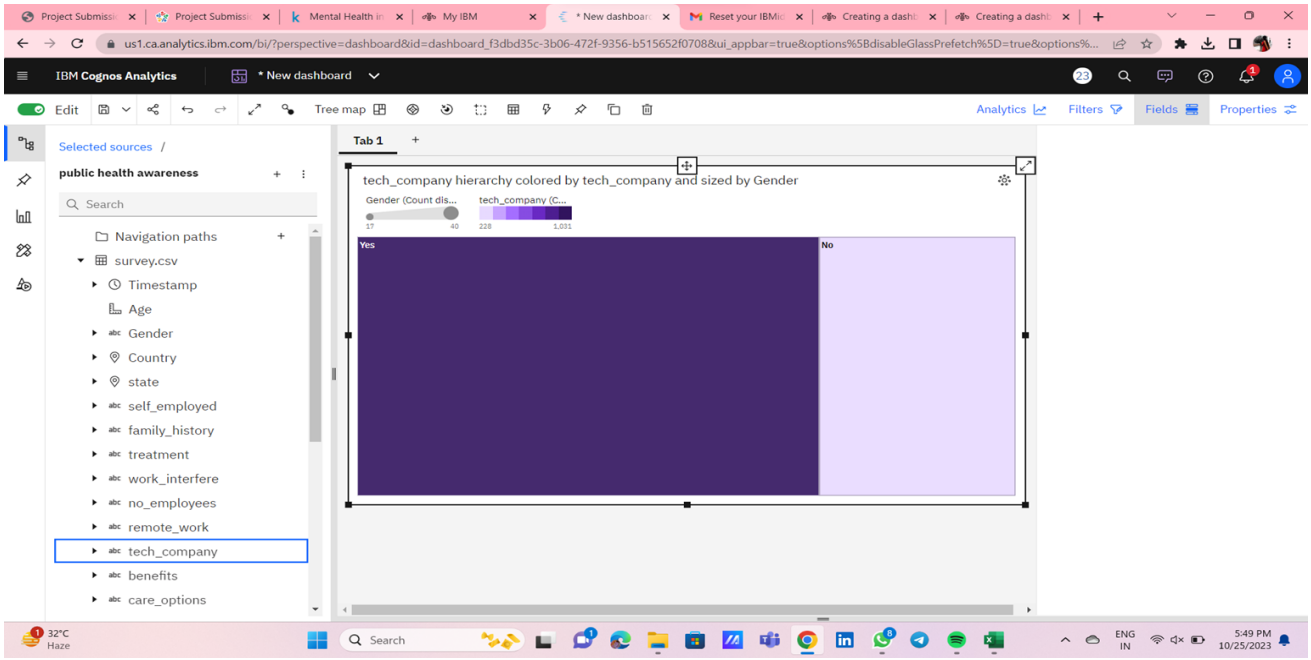


Fig 6.2

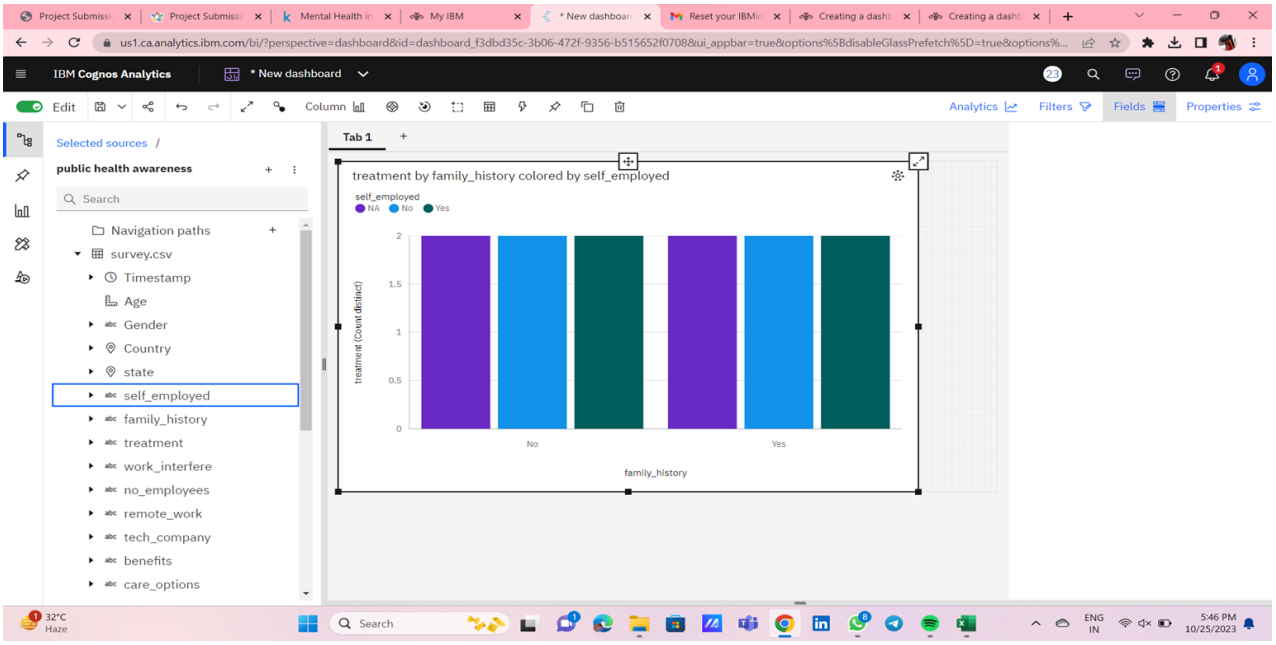


Fig 6.3

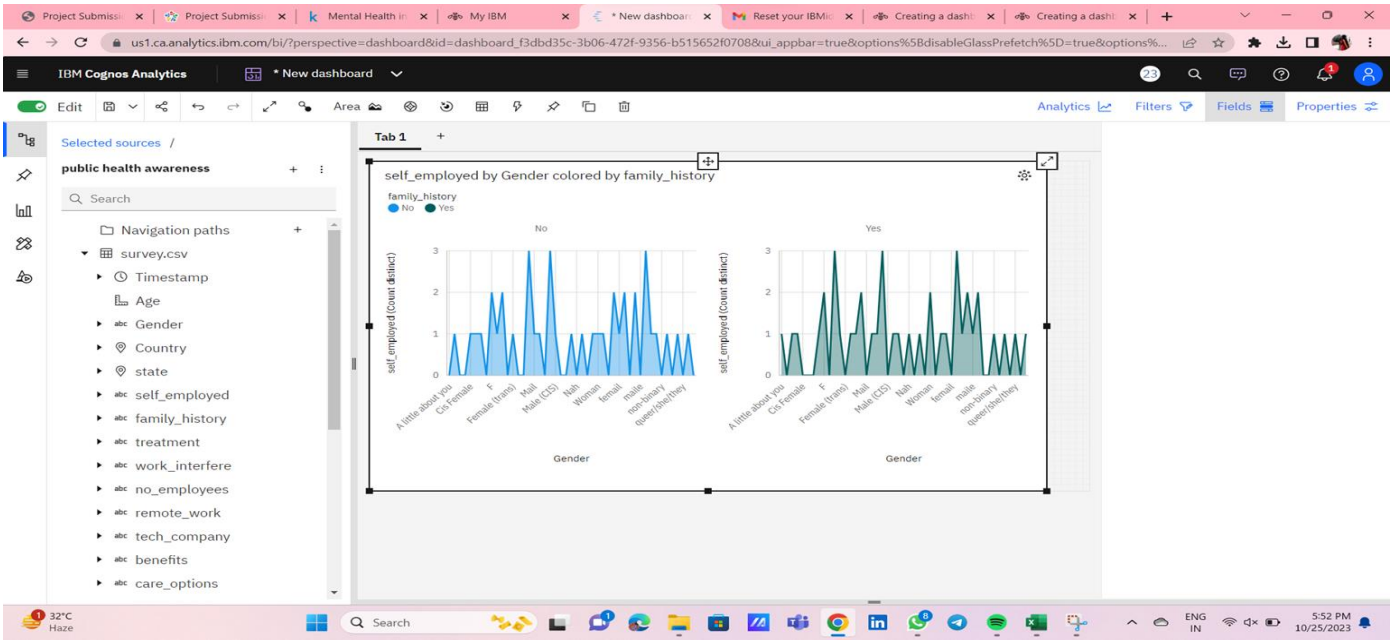
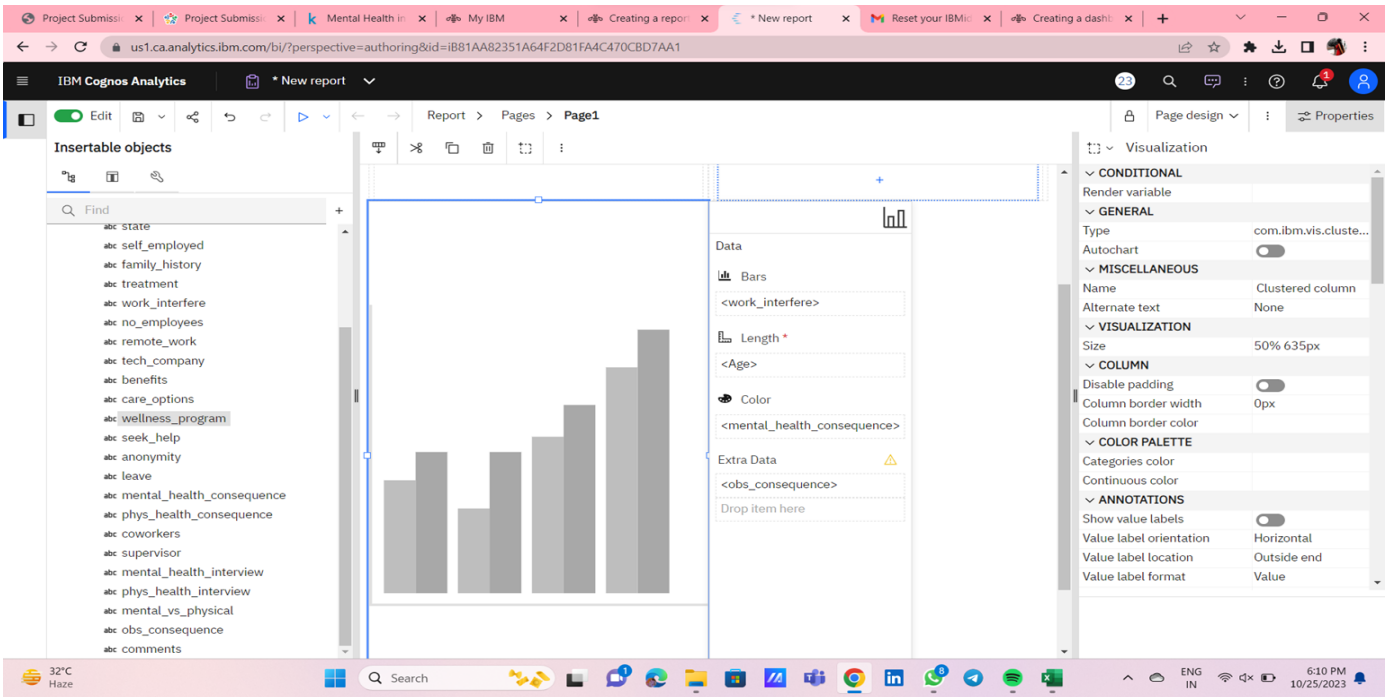

IBM Cognos Analytics
* New dashboard
Edit
Area
Analytics
Filters
Fields
Properties
Selected sources /
public health awareness
Search
Navigation paths
survey.csv
Timestamp
Age
Gender
Country
state
self_employed
family_history
treatment
work_interfere
no_employees
remote_work
tech_company
benefits
care_options
Tab 1
self_employed by Gender colored by family_history
family_history
No
Yes
self_employed (Count distinct)
Gender

Fig 6.4

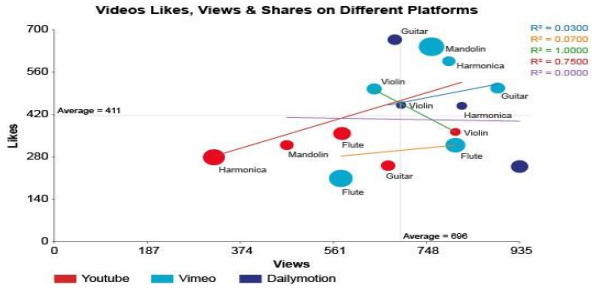

IBM Cognos Analytics
* New report
Edit
Report > Pages > Page1
Page design
Properties
Insertable objects
Find
state
self_employed
family_history
treatment
work_interfere
no_employees
remote_work
tech_company
benefits
care_options
wellness_program
seek_help
anonymity
leave
mental_health_consequence
phys_health_consequence
coworkers
supervisor
mental_health_interview
phys_health_interview
mental_vs_physical
obs_consequence
comments
Data
Bars
<work_interfere>
Length *
<Age>
Color
<mental_health_consequence>
Extra Data
<obs_consequence>
Drop item here
Visualization
CONDITIONAL
Render variable
GENERAL
Type
com.ibm.vis.cluste...
Autochart
MISCELLANEOUS
Name
Clustered column
Alternate text
None
VISUALIZATION
Size
50% 635px
COLUMN
Disable padding
Column border width
0px
Column border color
COLOR PALETTE
Categories color
Continuous color
ANNOTATIONS
Show value labels
Value label orientation
Horizontal
Value label location
Outside end
Value label format
Value

CHAPTER 6

PROPOSED INNOVATION TECHNIQUE

Exploratory Data Analysis (EDA) involves a variety of techniques and methods to gain insights into your data. Here are some common techniques used in EDA:

1. **Descriptive Statistics**:This involves calculating summary statistics such as mean, median, standard deviation, and percentiles. These statistics provide a basic understanding of the central tendency and spread of your data.



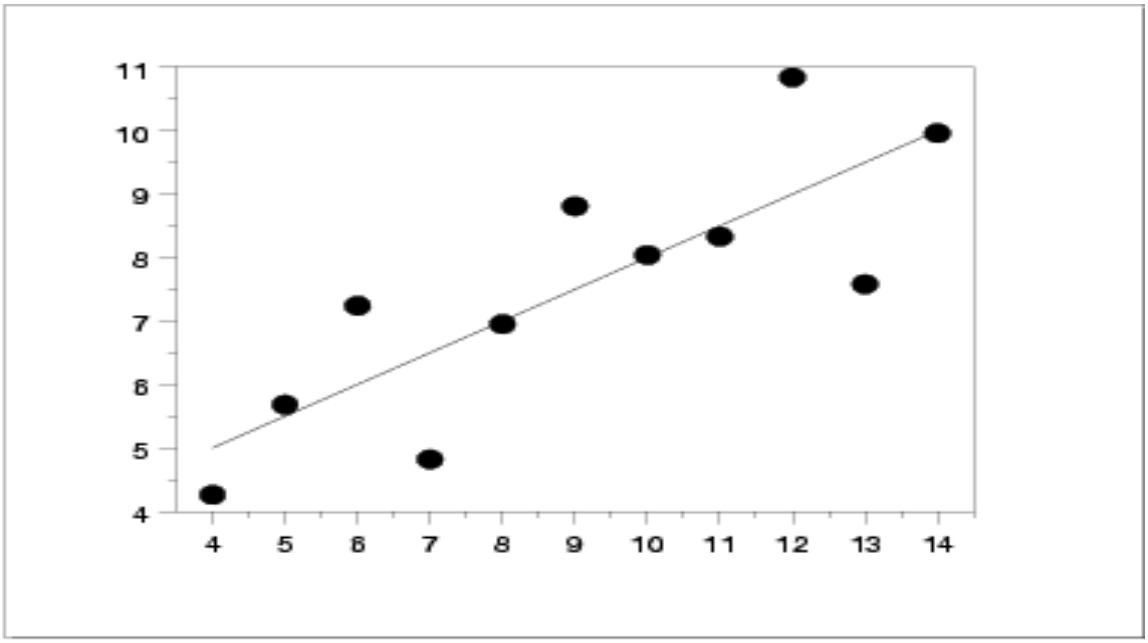
2. **Data Visualization:** Data visualization is a powerful EDA technique. It includes creating various plots and charts to visualize data distribution, relationships, and patterns. Common visualization techniques include:

- Histograms: To show the distribution of a single variable.
- Box plots: To display the distribution of a variable and identify outliers.
- Scatter plots: To visualize relationships between two variables.
- Bar charts: To compare categories or groups.

3. **Correlation Analysis:** This technique is used to measure the strength and direction of relationships between variables. Common measures of correlation include Pearson correlation coefficient for linear relationships and Spearman rank correlation for non-linear relationships.

4. **Data Cleaning:** Cleaning the data is a crucial EDA step. This involves handling missing data, dealing with outliers, and removing duplicates to ensure the quality of the data.

5. **Exploring Categorical Data:** For categorical variables, you can use techniques like frequency tables, bar charts, and pie charts to understand the distribution of categories.



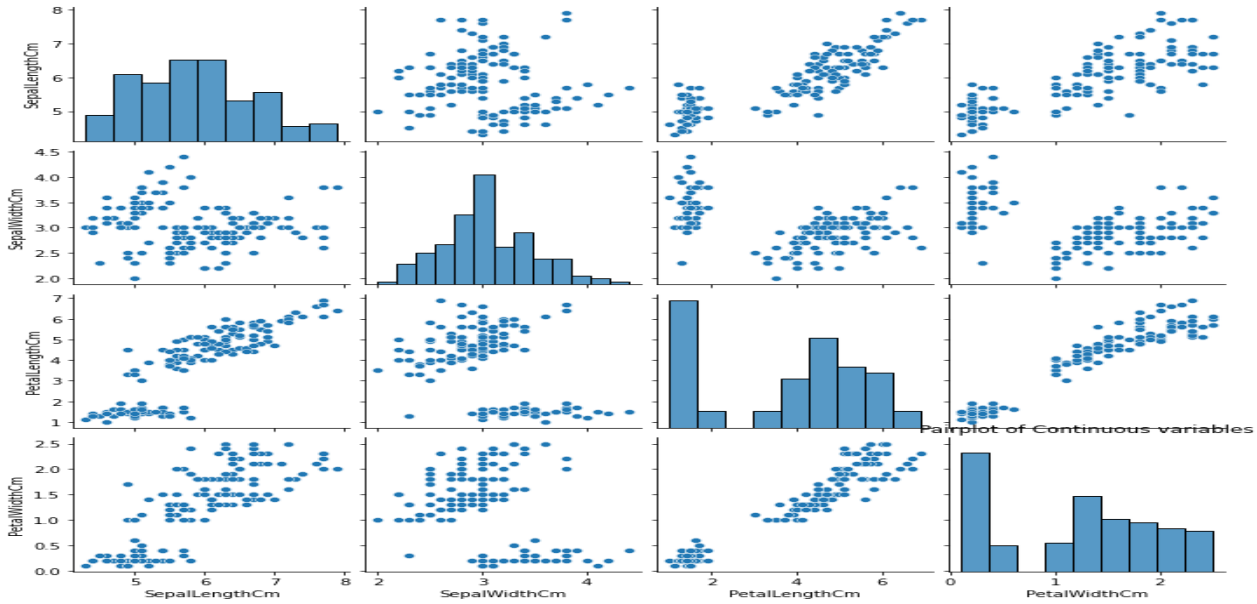
6. **Dimensionality Reduction:** If you have a high-dimensional dataset, you may use dimensionality reduction techniques such as Principal Component Analysis (PCA) to reduce the number of features while retaining important information.

7. **Outlier Detection:** Identifying and handling outliers can be essential in EDA. Common techniques include the use of box plots, z-scores, or the IQR (Interquartile Range) method.

8. **Feature Engineering:** EDA can help you discover new features or transformations of existing features that may be more informative for your analysis or modeling.

9. **Time Series Analysis**:If your data is time-series data, time series analysis

techniques, such as autocorrelation and moving averages, can be employed to explore temporal patterns.



10. **Hypothesis Testing**: EDA can involve preliminary hypothesis testing to assess the significance of observed differences or relationships in the data.

11. **Geospatial Analysis**:For datasets with geographic information, geospatial EDA techniques can be used to create maps and visualize spatial patterns.

12.**Clustering and Segmentation**: Clustering techniques, such as K-means clustering, can help uncover natural groupings within the data.

13. **Text Analysis**: For text data, techniques like word frequency analysis and sentiment analysis can provide insights.

14. **Advanced Visualization**: Techniques like heatmaps, violin plots, and pair plots (for multivariate data) can offer more advanced visualization options.

EDA is not a one-size-fits-all process, and the choice of techniques depends on the nature of the data and the specific objectives of the analysis.

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

CONCLUSION

Telecom customer churn prediction is a crucial aspect of the telecommunications industry, with substantial financial implications and the potential to enhance customer satisfaction. In this conclusion, we summarize the key takeaways and insights gained from this predictive analysis:

FUTURE SCOPE:

1. **Advanced Machine Learning and AI**:As machine learning and artificial intelligence (AI) technologies continue to advance, customer churn prediction models will become more accurate and sophisticated. Deep learning techniques, ensemble models, and natural language processing (NLP) can be applied to analyze a wider range of data sources, including unstructured data like customer reviews and social media posts.

2. **Big Data Integration**:With the increasing volume of data generated by businesses and customers, the integration of big data technologies will play a vital role in customer churn prediction. Businesses will need to leverage data lakes and cloud computing platforms to process and analyze vast amounts of data efficiently.

3. **Real-time Churn Prediction:**Real-time or near-real-time churn prediction will become more prevalent. Companies will use streaming data and event processing to identify potential churn indicators as they happen, allowing for proactive intervention to retain customers.

4. **Personalization and Customer Segmentation**:Customer churn models will focus on personalization by identifying unique reasons for churn for each customer and tailoring retention strategies accordingly. Advanced segmentation techniques will help businesses understand the distinct needs and preferences of various customer segments.

5. **Predictive Analytics Ecosystem**: A more comprehensive predictive analytics ecosystem will evolve, where customer churn prediction is integrated with other predictive models for cross-functional insights. This will enable businesses to optimize decision-making across departments such as marketing, sales.