Interactive Churn Analytics: From SQL to Dash Plotly - Integrating Data Management, Predictive Modeling, and Visualization for Churn Insights

Garima Choudhary

EC Utbildning

Kunskapskontroll2_ProjectDS

2024 October

Abstract

This project aims to analyze and predict customer churn using a combination of SQL, Power BI, Dash, and machine learning techniques. Customer churn, the loss of customers over time, is a critical concern for businesses. By leveraging historical customer data, this project performs data exploration and transformation in SQL to identify key patterns and trends in churn behavior. The transformed data is visualized using Power BI, providing interactive dashboards that highlight churn rates based on demographics, services, and customer behaviors.

In addition, a Random Forest algorithm is employed to build a machine learning model that predicts future customer churn. The model is trained on historical data and evaluated using classification metrics. The project concludes by applying the trained model to new customer data to forecast potential churners, enabling proactive business strategies to improve customer retention.

To enhance interactivity and accessibility, a Dash web app is integrated into the project, offering dynamic visualizations and real-time exploration of churn predictions. This web-based dashboard allows users to easily access key insights and predictions, empowering businesses to monitor churn patterns and make data-driven decisions to minimize churn and maintain customer loyalty.

Keywords

- Customer Churn
- SQL
- Power BI
- Data Exploration
- Dash
- Data Visualization
- Machine Learning
- Random Forest
- Churn Prediction
- Predictive Analytics
- Data Transformation
- Business Intelligence
- Customer Retention

Acknowledgement

I would like to express my gratitude towards Mr. Antonio Prgomet, my Mentor, for his continuous support and efforts. And I want to say thank you to everybody in my class for being so kind and generous.

Contents

Abstract 2
Keywords 3
Acknowledgement 4
1. Introduction 7
1.1 Data Exploration and Transformation (SQL) 8
1.2 Data Visualization (Power BI) 8
1.3 Churn Prediction (Machine Learning - Random Forest) 8
1.4 Interactive Dashboard (Dash) 8
2. Theory 10
2.1 Customer Churn Analysis 10
2.2 SQL for Data Exploration and Transformation 10
2.3 Power BI for Data Visualization 11
2.4 Dash for Interactive Web-Based Dashboards 11
2.5 Random Forest Algorithm for Churn Prediction 12
2.6 Evaluation Metrics 12
2.7 Feature Importance 12
3.Methodology 13
3.1 Data Collection 13
3.2 Data Preprocessing 14
3.3 Exploratory Data Analysis (EDA) 14
3.4 Model Building: Random Forest Algorithm 14
3.5 Model Evaluation 14
3.6 Making Predictions on New Data 15
3.7 Results and Insights 15
3.8 Data Visualization 15
3.9 Interactive Dashboard with Dash 15
4. Results and Discussion 16
4.1Confusion Matrix Results 17
True Negatives 17
False Positives 17
False Negatives 17
True Positives 17
4.2 Performance Metrics 17

4.3 Overall Accuracy: 84% 18
5. Conclusion 20
6.Självutvärdering 21
Appendix A 21
References 23

1.Introduction

Customer churn, or customer attrition, refers to the loss of customers over a given period. It is a key metric for businesses, particularly in sectors such as telecommunications, retail, and subscription-based services. High churn rates can negatively impact business profitability, as acquiring new customers is often more expensive than retaining existing ones. Understanding and predicting churn helps businesses address issues proactively, improve retention strategies, and enhance customer satisfaction.

In this project, we focus on analyzing customer churn using SQL, Power BI, and machine learning techniques. The primary goal is to explore historical customer data, identify key factors contributing to churn, and predict future churn events using a Random Forest model. The project consists of three main phases:

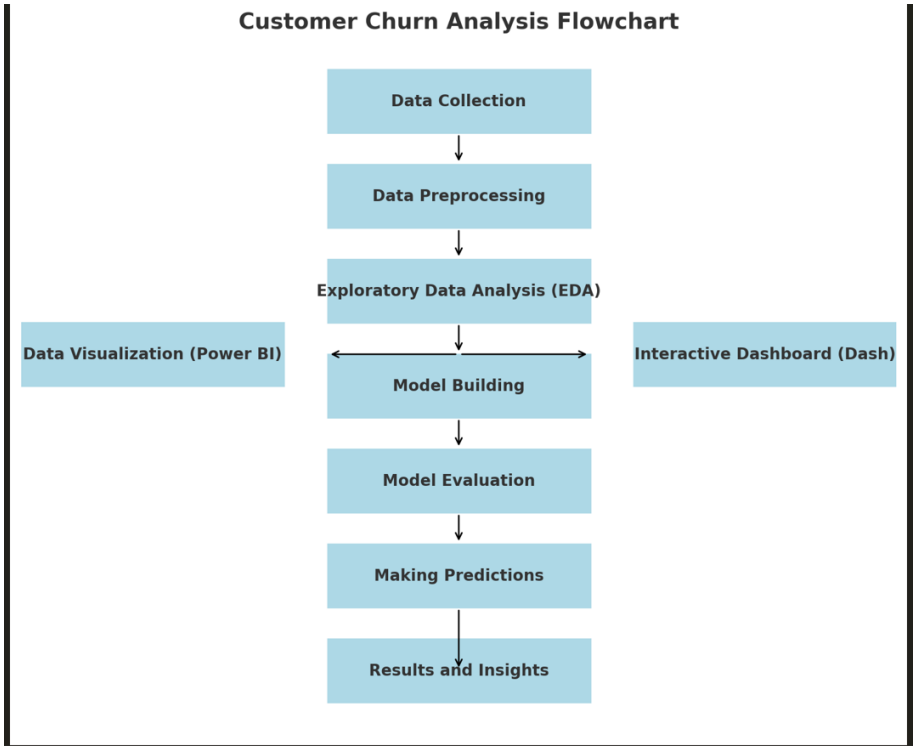

Image:1 Project Process Flow

1.1 Data Exploration and Transformation (SQL)

In this phase, we load, clean, and transform the raw customer data using SQL. Data exploration is performed to uncover patterns in customer demographics, services, and behavior that may influence churn. SQL queries are used to check for missing values, calculate churn rates, and identify key metrics.

1.2 Data Visualization (Power BI)

The second phase focuses on creating visualizations using Power BI. Interactive dashboards help businesses gain insights into churn trends based on age, marital status, payment methods, and service usage. These visualizations provide a clearer understanding of churn distribution and highlight areas for improvement.

1.3 Churn Prediction (Machine Learning- Random Forest)

In the final phase, a Random Forest algorithm is employed to predict customer churn based on historical data. The model is evaluated for accuracy and used to predict future churn, enabling businesses to implement targeted retention strategies.

By the end of this project, businesses will have a clearer understanding of the factors driving churn and a predictive model to help reduce customer attrition and improve retention.

1.4 Interactive Dashboard (Dash)

The final phase integrates **Dash**, a web-based Python framework, to create an interactive dashboard. This allows users to explore churn predictions and insights in real time through a web interface. Unlike static reports, Dash offers dynamic filtering and visualization options, providing decision-makers with the ability to interact with the data, track key metrics, and explore scenarios. This enhances the accessibility of the results and allows businesses to respond quickly to emerging churn trends.

By the end of this project, businesses will have a clearer understanding of the factors driving churn, an interactive dashboard for ongoing analysis, and a predictive model to help reduce customer attrition and improve retention.

2. Theory

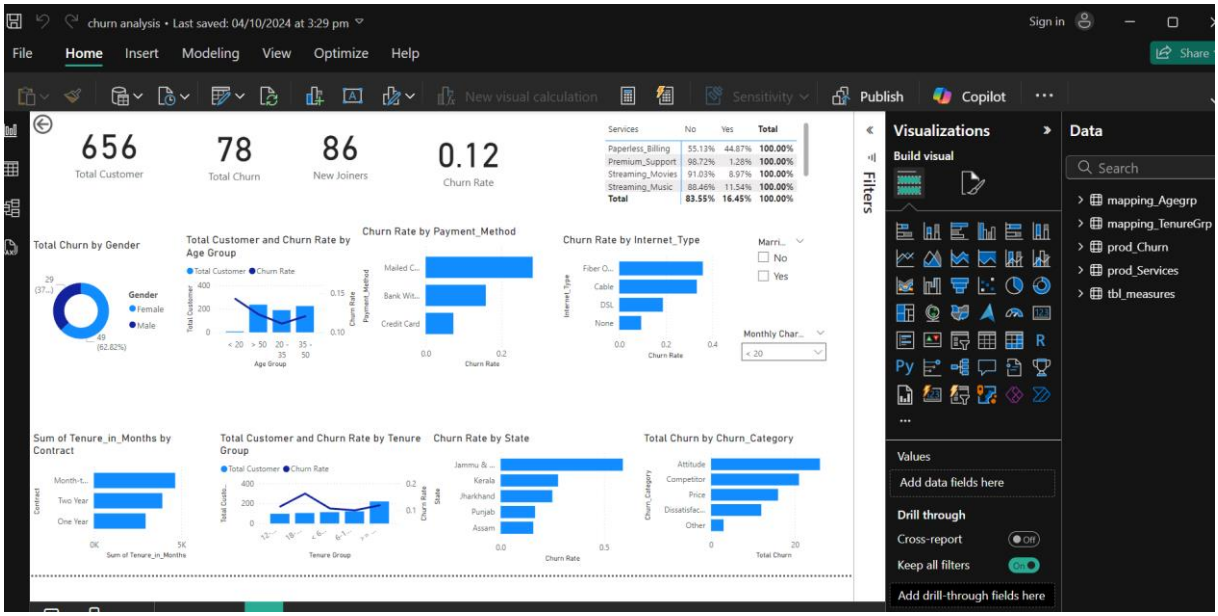

Image:2 Highlights of Project Outcomes

2.1 Customer Churn Analysis

Customer churn refers to the rate at which customers stop using a service. A high churn rate can hurt business profitability. Factors such as product quality, customer service, and pricing influence churn. Understanding these factors allows businesses to implement strategies to reduce churn.

2.2 SQL for Data Exploration and Transformation

SQL is a powerful language used for managing relational databases. In this project, SQL is utilized to:

Clean and transform data by removing duplicates, handling missing values, and structuring the data for analysis.

Explore data by querying customer demographics, contract types, and service usage to understand their impact on churn.

Aggregate and group data to calculate churn rates by attributes like age, state, and service type.

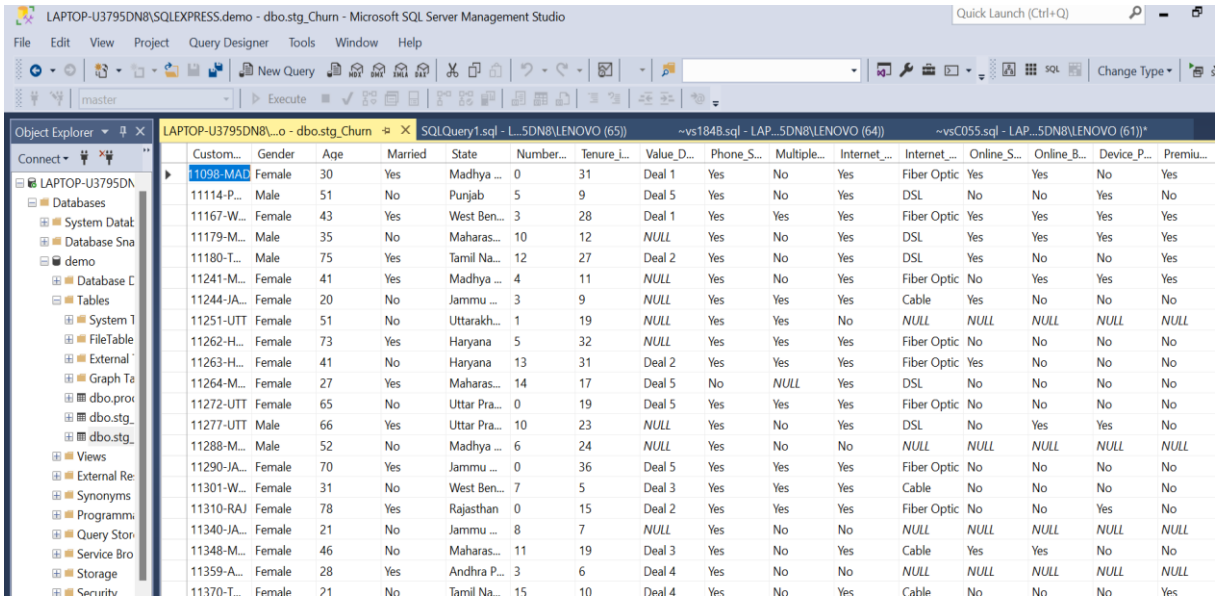

Image: 3 Visualization of Information

2.3 Power BI for Data Visualization

Power BI is used for:

Data visualization through charts and graphs to understand the distribution of churn across customer segments.

Exploratory data analysis (EDA) via interactive dashboards that enable businesses to explore customer demographics in relation to churn.

Churn insights reports, highlighting churn rates by various customer attributes.

2.4 Dash for Interactive Web-Based Dashboards

Dash is a Python framework used to create interactive web applications and data visualizations. In this project, Dash enhances data accessibility by:

- Offering an interactive, web-based dashboard that enables users to explore customer churn predictions in real-time.
- Allowing decision-makers to filter, sort, and drill down into the data dynamically, without needing to rerun SQL queries or refresh Power BI reports.

- Providing a user-friendly interface to display machine learning predictions, feature importance, and key customer metrics. This web app allows users to continuously monitor churn rates and make proactive adjustments to customer retention strategies.
  Dash complements static visualizations by providing a highly interactive tool for deeper insights into churn patterns and trends.

2.5 Random Forest Algorithm for Churn Prediction

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy. Key components include:

Decision Trees: Subsets of data are split based on features like age and contract type to make predictions.

Ensemble Learning: Multiple decision trees are combined to reduce overfitting and improve robustness.

Feature Importance: Random Forest provides insight into which features most influence churn predictions.

2.6 Evaluation Metrics

The model is evaluated using:

Confusion Matrix to assess prediction accuracy by comparing actual vs. predicted results.

Precision and Recall to measure the model's effectiveness in identifying churned customers.

F1-Score to balance precision and recall.

2.7 Feature Importance

Feature importance provides insight into which customer attributes (such as monthly charges, contract type, and service usage) are most relevant for predicting churn.

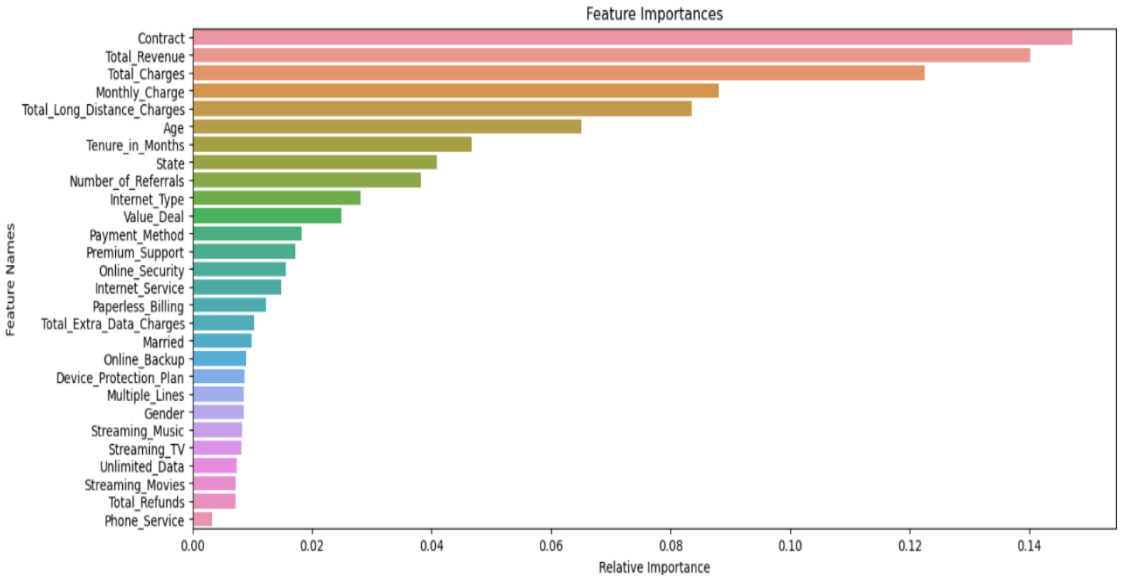

Image:4 Feature Importance

3.Methodology

3.1 Data Collection

The dataset for this project was sourced from SQL Server and stored in Excel files with customer demographics, contract details, and financial information. The data was extracted from the staging table stg_Churn.

3.2 Data Preprocessing

Data preprocessing included:

Handling Missing Values: Missing values in columns like Value_Deal and Internet_Type were filled with default values.

Removing Unnecessary Columns: Irrelevant columns, such as Customer_ID, were removed.

Encoding Categorical Variables: Categorical data (e.g., Gender, State, Contract) was converted into numerical values for the Random Forest model.

Target Variable Encoding: The Customer_Status column was mapped into binary values (0 for "Stayed" and 1 for "Churned").

3.3 Exploratory Data Analysis (EDA)

SQL queries were used to explore relationships between features such as age, contract type, and service usage with churn. This analysis informed the selection of features for model building.

3.4 Model Building: Random Forest Algorithm

Key steps in model building included:

Feature Selection based on EDA findings.

Data Splitting into training (80%) and test (20%) sets.

Model Training using 100 decision trees (n_estimators=100) to predict churn.

3.5 Model Evaluation

The model was evaluated using:

Confusion Matrix to track true positives, false positives, true negatives, and false negatives.

Classification Report to calculate precision, recall, and F1-score for each class (Stayed and Churned).

Feature Importance Analysis to visualize which factors (e.g., monthly charges, contract type) influenced churn predictions.

3.6 Making Predictions on New Data

New customer data was preprocessed similarly to training data, and predictions were made using the trained Random Forest model. The predictions were saved in a CSV file for business decision-making.

3.7 Results and Insights

Model Performance

The model achieved 84% accuracy in predicting customer churn. While it performed well for predicting Stayed customers, its recall for Churned customers was 65%, indicating room for improvement in identifying potential churn.

Feature Importance

Monthly charges and contract type were the most influential factors in predicting churn, highlighting areas where retention strategies can be targeted.3.8 Data Visualization

Power BI dashboards were created to visualize churn rates by customer attributes, while a feature importance plot in Jupyter Notebook identified key drivers behind churn.

3.9 Interactive Dashboard with Dash

In addition to Power BI, **Dash** was used to develop a web-based interactive dashboard for real-time exploration of customer churn predictions. This dashboard allows users to dynamically filter data, interact with predictive insights, and drill down into specific features such as contract type, monthly charges, or service usage to understand their contribution to churn. Dash provides a more flexible and accessible platform for business decision-makers, enabling them to continuously monitor and adjust customer retention strategies based on real-time data analysis.

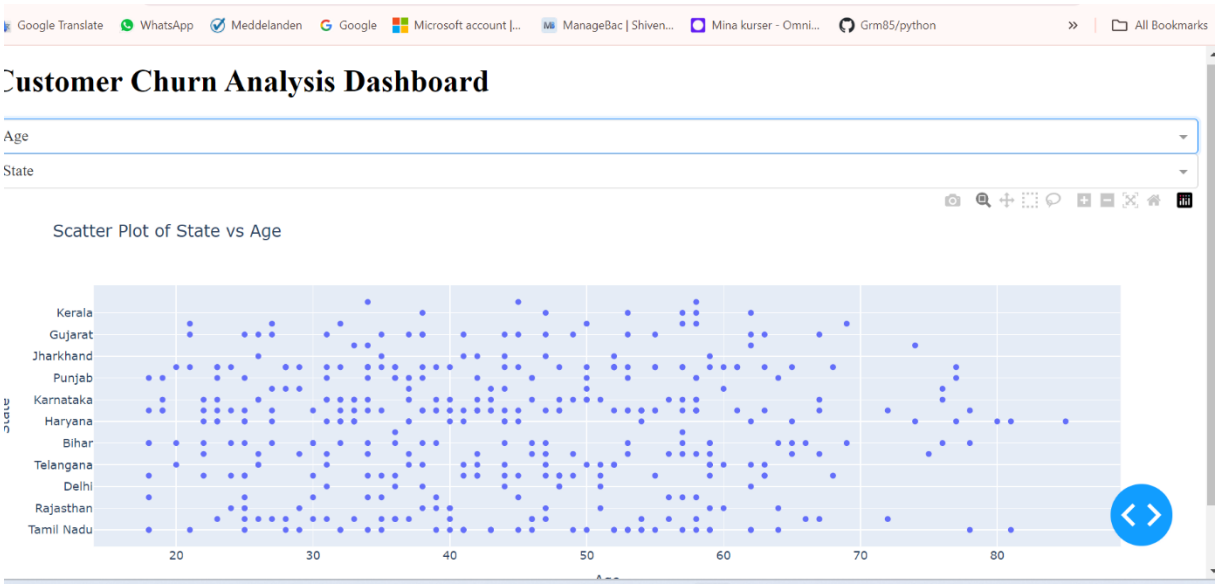

Image:5 Screenshot of the Dash Web Application Interface

4. Results and Discussion

In predictive modeling, evaluating the performance of a classifier involves assessing how well the model distinguishes between different classes (in this case, customers who churned vs. those who stayed). The **Random Forest Classifier** produced the following outcomes:

```
[[783  64]
 [126 229]]

Classification Report:
              precision    recall  f1-score   support

           0       0.86      0.92      0.89       847
           1       0.78      0.65      0.71       355

    accuracy                           0.84      1202
   macro avg       0.82      0.78      0.80      1202
weighted avg       0.84      0.84      0.84      1202
```

Image:6 Results Overview: Customer Churn Predictions

4.1Confusion Matrix Results

True Negatives (TN): 783

These represent customers who did not churn, and the model correctly predicted they would stay.

False Positives (FP): 64

These are customers predicted to churn but who actually stayed. This misclassification results in overestimating churn, potentially leading to unnecessary retention efforts.

False Negatives (FN): 126

These are customers who churned but were predicted to stay. This is a significant type of error in churn analysis, as retaining such customers would be difficult without knowing they are at risk.

True Positives (TP): 229

These are customers who churned and were correctly identified by the model. Correctly classifying these customers enables targeted retention strategies.

4.2 Performance Metrics

The model's performance is evaluated using several key metrics, each with its own implications for business decisions:

Precision:

Stayed (0): 86%

Precision for non-churning customers (0) indicates that out of all customers predicted to stay, 86% actually stayed. This reflects the model's accuracy in not falsely labelling churners as non-churners.

Churned (1): 78%

Precision for churning customers (1) shows that 78% of those predicted to churn actually did. This is critical in preventing wasted efforts on customers who were unlikely to leave.

Recall:

Stayed (0): 92%

Recall for non-churners measures the proportion of actual non-churning customers who were correctly predicted to stay. A high recall for non-churners (92%) is desirable, ensuring that the model correctly identifies most loyal customers.

Churned (1): 65%

Recall for churners, however, stands at 65%. This means that 35% of actual churners were missed by the model. Improving recall for churners would help in identifying more customers at risk of leaving.

F1-Score:

Stayed (0): 0.89

Churned (1): 0.71

The F1-Score is a balanced metric between precision and recall, giving a single number to represent the trade-off between these two metrics. While the score is higher for non-churners (0.89), the relatively lower score for churners (0.71) highlights the need for improving the model's sensitivity to churners.

4.3 Overall Accuracy: 84%

Accuracy refers to the percentage of correct predictions out of all predictions made by the model. An accuracy of 84% shows that the model correctly predicted customer behavior in most cases, but it does not reflect the nuances of how well the model handles churners versus non-churners.

Interpretation and Implications

The model shows strong performance in predicting customers who will stay, with high precision (86%) and recall (92%). However, for customers likely to churn, the performance, though decent (78% precision, 65% recall), leaves room for improvement. Given the business objective of minimizing churn, enhancing the model's ability to identify potential churners—especially through increasing recall—should be a key focus for future iterations.

Furthermore, the relatively lower F1-Score for churners (0.71) suggests that the balance between precision and recall can be optimized, perhaps by adjusting the decision threshold or fine-tuning the model parameters.

In conclusion, while the Random Forest Classifier performs well overall, targeted improvements can make it more effective at catching potential churners, which is crucial for developing proactive retention strategies.

5. Conclusion

This project successfully developed a Random Forest model to predict customer churn with 84% accuracy. While the model performed well in predicting customers who stayed, there is room for improvement in identifying those who churned. By refining features and exploring other algorithms, the model's accuracy can be further enhanced, which could help businesses reduce churn and improve retention strategies.

A comprehensive approach was used, combining SQL for data exploration, Power BI for visual insights, and the Random Forest algorithm for prediction. SQL queries provided key insights into demographic factors like age, marital status, and service usage, which were then visualized in Power BI to help communicate findings effectively.

Additionally, an interactive dashboard built with **Dash** enhances the accessibility of the predictions, allowing decision-makers to explore churn trends in real time. The Random Forest model identified patterns in customer behavior, was validated against historical data, and its predictions were exported for business use. This integration of tools provided a holistic view of churn, generating actionable insights for improving customer engagement and retention efforts.

6.Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

Svar: Funderar på arbetsflödet och hur man kan förbättra det.

2. Vilket betyg du anser att du skall ha och varför.

Svar: Enligt mitt arbete.

3. Något du vill lyfta fram till Antonio?

Svar: Nej.

Appendix A

Image 1 Project Process Flow…….
Image 2 Highlights of Project Outcomes……
Image 3 Visualization of Information……
Image 4 Feature Importance…….
Image 5 Screenshot of the Dash Web Application Interface…….
Image 6 Results Overview: Customer Churn Predictions…….

References

GeeksforGeeks. https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

W3Schools.com. (n.d.). https://www.w3schools.com/sql/sql_constraints.asp

Mihart. (2024, August 6). *Basic concepts for designers in the Power BI service - Power BI*. Microsoft Learn. https://learn.microsoft.com/en-us/power-bi/fundamentals/service-basic-concepts

DASH Documentation & User Guide | Plotly. (n.d.). https://dash.plotly.com/

W3Schools.com. (n.d.-b). https://www.w3schools.com/sql/sql_ref_mysql.asp