**VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY**

**UNIVERSITY OF INFORMATION TECHNOLOGY**

**FACULTY OF INFORMATION SYSTEMS**

**\*\*\*\*\*\***

****

**Final Project Report**

**TELCO CUSTOMER CHURN ETL, OLAP, AND DATA MINING SOLUTION**

**Lecturer:**MSc. Nguyen Thi Kim Phung

**Course:** Decision Support & Business Intelligence

**Class:** MSIS4263.P21.CTTT

**Student:** Ngo Manh Khuong – 21522241

Ngo Cong Huan - 21520881

Ho Chi Minh city, June 2025

**TABLE OF CONTENTS**

[**I.** **SSIS process** 3](#_Toc200295924)

[**II.** **Analyzing and reporting processes** 11](#_Toc200295925)

[**a)** **Describe the schema (star/snow)** 11](#_Toc200295926)

[**b)** **Process of building a cube** 12](#_Toc200295927)

[**c)** **Using analysis service (manual), creating reports (Power BI), analysis service (MDX) and Pivot table (Excel)** 13](#_Toc200295928)

[**III.** **Data mining (2 algorithms, deep learning)** 34](#_Toc200295929)

1. **SSIS process**

**Kaggle dataset**: <https://www.kaggle.com/code/anubhavgoyal10/customer-churn-prediction-eda-ann/input>

Table 1: Dataset information

| **Column Name** | **Description** |
| --- | --- |
| customerID | Unique identifier for each customer |
| gender | Customer's gender (Male, Female) |
| SeniorCitizen | Indicates if the customer is a senior (1 = Yes, 0 = No) |
| Partner | Whether the customer has a partner (Yes/No) |
| Dependents | Whether the customer has dependents (Yes/No) |
| tenure | Number of months the customer has stayed |
| PhoneService | Whether the customer has phone service (Yes/No) |
| MultipleLines | Whether the customer has multiple lines (Yes, No, No phone service) |
| InternetService | Type of internet (DSL, Fiber optic, No) |
| OnlineSecurity | Internet security service (Yes/No) |
| OnlineBackup | Online backup service (Yes/No) |
| DeviceProtection | Device protection plan (Yes/No) |
| TechSupport | Technical support plan (Yes/No) |
| StreamingTV | Streaming TV service (Yes/No) |
| StreamingMovies | Streaming movie service (Yes/No) |
| Contract | Contract type (Month-to-month, One year, Two year) |
| PaperlessBilling | Whether the customer uses paperless billing (Yes/No) |
| PaymentMethod | Method of payment (e.g., Mailed check, Electronic check, Bank transfer) |
| MonthlyCharges | Current monthly charges |
| TotalCharges | Total amount charged to the customer |
| Churn | Target variable: whether the customer has churned (Yes/No) |

**Descriptions of Dimensions:**

**+ Dim\_Customer**

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| customerID | nvarchar(255) | Unique identifier for each customer (acts as primary key). |
| gender | nvarchar(255) | Gender of the customer (Male, Female). |
| SeniorCitizen | float | Indicates if the customer is a senior citizen (e.g., 0 = No, 1 = Yes). |
| Partner | nvarchar(255) | Whether the customer has a partner (Yes/No). |
| Dependents | nvarchar(255) | Whether the customer has dependents (Yes/No). |
| PaperlessBilling | nvarchar(255) | Whether the customer uses paperless billing (Yes/No). |

**Purpose:** Stores demographic and billing behavior details of customers.  
**BI Contribution:**

* Enables segmentation of churn by gender, senior status, partnership status, and billing preferences.
* Answers questions like:
  + *Are senior citizens more likely to churn?*
  + *Do customers with dependents stay longer?*
  + *Is paperless billing associated with lower churn?*

**+ Dim\_InternetService**

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| InternetServiceID | int | Surrogate key for the internet service dimension. |
| InternetService | nvarchar(255) | Type of internet service (DSL, Fiber optic, No...). |
| InternetAvailability | nvarchar(255) | General category: Has Internet, No Internet. |

**Purpose:** Describes the type and availability of internet service used.  
**BI Contribution:**

* Help identify churn patterns based on service availability.
* Distinguishes between users with No Internet vs. DSL or Fiber.
* Answers questions like:
  + *Does fiber optic service lead to better customer retention?*
  + *Are customers without internet more likely to leave?*

**+ Dim\_PaymentMethod**

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| PaymentMethodID | int | Surrogate key for the payment method. |
| PaymentCategory | nvarchar(255) | Category of payment method: Automatic or Manual. |
| PaymentMethod | nvarchar(255) | Specific payment method (Electronic check, Bank transfer, etc.). |

**Purpose:** Stores the specific and grouped methods of customer payments.  
**BI Contribution:**

* Enables analysis of how payment habits affect churn.
* Splits into Automatic vs. Manual, helping understand convenience vs. churn.
* Answers questions like:
  + *Do automatic payments reduce churn rates?*
  + *Which payment methods are common among loyal customers?*

**+ Dim\_AdditionalServices**

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| AdditionalServicesID | int | Surrogate key. |
| PhoneService | nvarchar(255) | Whether the customer has phone service. |
| MultipleLines | nvarchar(255) | Multiple phone lines (Yes/No). |
| OnlineSecurity | nvarchar(255) | Whether online security is enabled. |
| OnlineBackup | nvarchar(255) | Whether online backup is enabled. |
| DeviceProtection | nvarchar(255) | Whether device protection is enabled. |
| TechSupport | nvarchar(255) | Whether tech support is included. |
| StreamingTV | nvarchar(255) | Whether streaming TV is enabled. |
| StreamingMovies | nvarchar(255) | Whether streaming movies are enabled. |

**Purpose**: Captures whether customers subscribed to value-added services (phone, streaming, backup, etc.).  
**BI Contribution**:

* Helps identify product bundles that retain customers longer.
* Correlates specific services (e.g., streamingTV) with loyalty.
* Answers questions like:
  + *Does having tech support or online security reduce churn?*
  + *Which service combinations have the highest retention?*

**+ Dim\_Contract**

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| contractID | int | Surrogate key. |
| ContractGroup | nvarchar(255) | Grouping (Short-term, Long-term). |
| Contract | nvarchar(255) | Specific contract type (Month-to-month, One year, Two year). |

**Purpose**: Contains contract types and their length classification.  
**BI Contribution**:

* Shows how contract commitment level affects churn.
* Groups customers by short-term vs. long-term behavior.
* Answers questions like:
  + *Do month-to-month customers churn more than yearly ones?*
  + *Is long-term commitment effective in reducing churn?*

**+ Dim\_Tenure**

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| tenureID | int | Surrogate key. |
| tenureRange | nvarchar(255) | Bucketed tenure range (e.g., 0–12 months, 13–24 months, etc.). |
| tenureGroup | nvarchar(255) | Grouping of tenure: Short-term, Medium-term, Long-term. |
| tenure | int | Actual number of months the customer has stayed. |

**Purpose**: Captures how long a customer has been with the company, both as raw months and grouped buckets.  
**BI Contribution**:

* Crucial for understanding customer lifecycle.
* Helps identify tenure-based retention strategies.
* Answers questions like:
  + *Are newer customers more likely to churn?*
  + *Which tenure group is most loyal or most at risk?*

**+ Dim\_Churn**

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| ChurnID | int | Surrogate key. |
| Churn | nvarchar(255) | Indicates whether the customer churned (Yes, No). |

**Purpose**: Indicates whether a customer has churned.  
**BI Contribution**:

* Used as a label or target metric in churn reporting, KPIs, and predictive analytics.
* Enables side-by-side comparisons of churned vs. retained groups.
* Answers questions like:
  + *What percentage of customers churned last month?*
  + *What are the profiles of customers who tend to churn?*

**+ CustomerChurnFact (Fact Table)**

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| FactID | int (PK) | Surrogate primary key for the fact table. |
| customerID | nvarchar(255) | Foreign key to Dim\_Customer. |
| InternetServiceID | int | Foreign key to Dim\_InternetService. |
| PaymentMethodID | int | Foreign key to Dim\_PaymentMethod. |
| AdditionalServicesID | int | Foreign key to Dim\_AdditionalServices. |
| contractID | int | Foreign key to Dim\_Contract. |
| tenureID | int | Foreign key to Dim\_Tenure. |
| ChurnID | int | Foreign key to Dim\_Churn. |
| MonthlyCharges | float | The amount charged to the customer monthly. |
| TotalCharges | float | The total amount charged to the customer. |

**Describe the SSIS process**

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1: Control Flow structure

In the first SQL task execution, this task deletes all data from the tables and reset the IDs for selected table.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 2: First SQL task execution

In the second SQL task execution, we add constraints for all tables.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 3: Second SQL task execution

In the last SQL task execution, we grouped the data for each table.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 4: Last SQL task execution

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5: First Data Flow Task

A screenshot of a computer

AI-generated content may be incorrect.

Figure 6: Second Data Flow Task

1. **Analyzing and reporting processes**
2. **Describe the schema (star/snow)**

**Schema Type: Star Schema**

* There is one central fact table: CustomerChurnFact.
* Surrounding it are several dimension tables:
  + Dim\_Customer
  + Dim\_InternetService
  + Dim\_PaymentMethod
  + Dim\_AdditionalServices
  + Dim\_Contract
  + Dim\_Tenure
  + Dim\_Churn
* All dimension tables are directly linked to the fact table via foreign keys (no hierarchies or snowflaked sub-dimensions).
* There is no further normalization of dimensions into sub-dimensions (which would be typical of a snowflake schema).

1. **Process of building a cube**

A screenshot of a computer menu

AI-generated content may be incorrect.

Figure 7: Structure of SSAS

A screenshot of a computer

AI-generated content may be incorrect.

Figure 8: Data Source View

A screenshot of a computer

AI-generated content may be incorrect.

Figure 9: Cube

1. **Using analysis service (manual), creating reports (Power BI), analysis service (MDX) and Pivot table (Excel)**

- Below are 10 Business intelligence scenarios:

**1. TenureGroup Customer Counts**

**Row (TenureGroup), Column (Churn), Values (Count of CustomerID)**

The retention team is trying to identify new subscribers (Short-term tenure) vs. long-term loyalists. By examining churn by tenure group, they can launch loyalty campaigns to retain those most likely to churn early.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 10: SSAS

A screenshot of a computer

AI-generated content may be incorrect.

Figure 11: MDX

A graph on a computer screen

AI-generated content may be incorrect.

Figure 12: PowerBI

A screenshot of a graph

AI-generated content may be incorrect.

Figure 13: Excel

**2. InternetService and ContractGroup Monthly Charges**

**Row (InternetService), Column (ContractGroup), Values (Average of MonthlyCharges)**

The pricing team is reviewing high-cost internet services and contract types (e.g., Long-term fiber internet users) to consider discounts or promotional offers next quarter.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 14: SSAS

A screenshot of a computer

AI-generated content may be incorrect.

Figure 15: MDX

A screenshot of a graph

AI-generated content may be incorrect.

Figure 16: PowerBI

A screenshot of a graph

AI-generated content may be incorrect.

Figure 17: Excel

**3. PaymentCategory and PaperlessBilling Total Charges**

**Row (PaymentCategory), Column (PaperlessBilling), Values (Sum of TotalCharges)**

The finance team wants to forecast total revenue and finds that customers using automatic payments and paperless billing generate more revenue. They will promote paperless billing to boost revenue.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 18: SSAS

A screenshot of a computer

AI-generated content may be incorrect.

Figure 19: MDX

A screenshot of a computer

AI-generated content may be incorrect.

Figure 20: PowerBI

A screenshot of a computer

AI-generated content may be incorrect.

Figure 21: Excel

**4. Contract and InternetAvailability Customer Counts**

**Row (Contract), Column (InternetAvailability), Values (Count of CustomerID)**

Customer service wants to know which contract types are popular among users with internet. This helps in targeting upgrades or bundles for contracts with frequent complaints.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 22: SSAS

A screenshot of a computer

AI-generated content may be incorrect.

Figure 23: MDX

A screenshot of a computer

AI-generated content may be incorrect.

Figure 24: PowerBI

A screenshot of a data

AI-generated content may be incorrect.

Figure 25: Excel

**5. TenureRange and SeniorCitizen Total Charges**

**Row (TenureRange), Column (SeniorCitizen), Values (Average of TotalCharges)**

Product development sees that senior citizens often have lower total charges and can now design affordable packages tailored for this demographic, especially newer customers.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 26: SSAS

A screenshot of a computer

AI-generated content may be incorrect.

Figure 27: MDX

A graph of blue bars

AI-generated content may be incorrect.

Figure 28: PowerBI

A screenshot of a computer

AI-generated content may be incorrect.

Figure 29: Excel

**6. InternetService and Dependents Monthly Charges**

**Row (InternetService), Column (Dependents), Values (Average of MonthlyCharges)**

Marketing identifies single customers without dependents paying high monthly fees for internet. These customers may churn, so a discount plan can help retain them.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 30: SSAS

A screenshot of a computer

AI-generated content may be incorrect.

Figure 31: MDX

A graph of blue squares

AI-generated content may be incorrect.

Figure 32: PowerBI

A screenshot of a data

AI-generated content may be incorrect.

Figure 33: Excel

**7. PaymentMethod Customer Counts**

**Row (PaymentMethod), Column (Churn), Values (Count of CustomerID)**

Operations want to find which payment methods are preferred or problematic. For example, if manual check users churn more, they can promote automatic payments via incentives.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 34: SSAS

A screenshot of a computer

AI-generated content may be incorrect.

Figure 35: MDX

A pie chart with numbers and a diagram

AI-generated content may be incorrect.

Figure 36: PowerBI

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 37: Excel

**8. ContractGroup and Gender Total Charges**

**Row (ContractGroup), Column (Gender), Values (Sum of TotalCharges)**

Sales examines gender-based revenue patterns to target high-revenue customers (e.g., females in long-term contracts) with exclusive promotions or perks.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 38: SSAS

A screenshot of a computer

AI-generated content may be incorrect.

Figure 39: MDX

A screenshot of a graph

AI-generated content may be incorrect.

Figure 40: PowerBI

A screenshot of a graph

AI-generated content may be incorrect.

Figure 41: Excel

**9. Tenure and Partner Monthly Charges**

**Row (Tenure), Column (Partner), Values (Average of MonthlyCharges)**

The analytics team explores if customers with partners and long tenure pay more. If so, they can create a family loyalty program that rewards staying longer as a couple.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 42: SSAS

A screenshot of a computer

AI-generated content may be incorrect.

Figure 43: MDX

A graph with blue lines

AI-generated content may be incorrect.

Figure 44: PowerBI

A graph on a white background

AI-generated content may be incorrect.

Figure 45: Excel

**10. InternetService and PaperlessBilling Customer Counts**

**Row (InternetService), Column (PaperlessBilling), Values (Count of CustomerID)**

Customer experience team checks how many customers per internet service use paperless billing. This helps with launching a sustainability or green billing campaign, promoting eco-friendly practices and boosting retention.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 46: SSAS

A screenshot of a computer

AI-generated content may be incorrect.

Figure 47: MDX

A graph with blue squares

AI-generated content may be incorrect.

Figure 48: PowerBI

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 49: Excel

**Table 2: Summary Table**

| **Scenario** | **BI Goal** | **Key Value** |
| --- | --- | --- |
| 1. TenureGroup x Churn | Retention campaign | Customer count |
| 2. InternetService x ContractGroup | Pricing discounts | Avg Monthly Charges |
| 3. PaymentCategory x PaperlessBilling | Revenue boost | Total Charges |
| 4. Contract x InternetAvailability | Upgrade strategy | Customer count |
| 5. TenureRange x SeniorCitizen | Senior package | Avg Total Charges |
| 6. InternetService x Dependents | Target singles | Avg Monthly Charges |
| 7. PaymentMethod x Churn | Incentivize auto-pay | Customer count |
| 8. ContractGroup x Gender | Gender-based promotion | Total Charges |
| 9. Tenure x Partner | Loyalty program | Avg Monthly Charges |
| 10. InternetService x PaperlessBilling | Paperless campaign | Customer count |

1. **Data mining (2 algorithms, deep learning)**

**Overview**: In the part, we will use 6 algorithms to predict Customer Churn in order to decide which model is the best for classification of Churn customer values. From the model, we will extract the rules for this dataset.

We will define which features are the most important to Churn values. We use Correlation matrix, Random Forest algorithm, Permutation on Test set to figure it out.

A screenshot of a graph

AI-generated content may be incorrect.

Figure 50: Correlation Matrix

A graph of a number of different types of data

AI-generated content may be incorrect.

Figure 51: Feature importance by Random Forest

A graph with blue and white bars

AI-generated content may be incorrect.

Figure 52: Permutation importance on Test set

After feature engineering, we can notice that Contract, Tenure, and Monthly Charges are 3 most important features affecting Churn outcomes.

A close-up of a graph

AI-generated content may be incorrect.

Figure 53: Confusion Matrixes

We use ROC Curves to evaluate all Models. The results show that Logistic Regression, Gradient Boosting, CatBoost are the highest accuracy models.

A graph showing the results of a logistic positive rate

AI-generated content may be incorrect.

Figure 54: ROC curves

After applying all models to predict the Churn values. We use Radar char to show important features of each model.

A group of colored hexagons

AI-generated content may be incorrect.

Figure 55: Top Feature radar charts

We print out the Churn values (yes = 1, no = 0) of the actual, Predicted **Logistic Regression**, and predicted **Gradient Boosting** values to compare it together.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 56: Predicted values comparison

Lastly, we define Rules for churn classification. The result below is the rule with Scaled values.

A screenshot of a computer program

AI-generated content may be incorrect.

Figure 57: Classification Rules for Churn (Scaled)

We transform again the rules from scaled values to original values which are more meaningful in Business intelligence decision.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 58: Classification Rules for Churn (Original Values)

**Explanation of Classification Rules**

**1. Contract is Month-to-month:**

* The rule only applies to customers with a Month-to-month contract. These customers are known to be more prone to churn compared to long-term contract holders.

**2. If MonthlyCharges <= 68.63:**

Then check tenure:

* If tenure <= 3 → **No Churn**

Customers with low charges and very new are predicted not to churn.

* If tenure <= 14 → **Churn**

If the customer is moderately new (3 to 14 months), the model predicts churn.

Else → **No Churn**

If tenure is longer than 14 months, churn risk drops again.

**3. If MonthlyCharges <= 93.68 (and more than 68.63):**

Regardless of tenure, if **Contract is Month-to-month or One year**:

→ **No Churn**

These customers pay moderately high charges but are still likely to stay.

**4. If MonthlyCharges > 93.68:**

Again, if Contract is Month-to-month or One year:

→ **No Churn**

Even high-paying customers are predicted not to churn, possibly because they're receiving more services and are satisfied.

We show the Decision tree for Churn classification.

A diagram of a decision tree

AI-generated content may be incorrect.

Figure 59: Churn decision tree

We also add a Grouped bar char about Average Monthly Charges by Tenure bins and Churn values.

A graph of blue and orange bars

AI-generated content may be incorrect.

Figure 60: Average Monthly Charges by Tenure Bin and Churn

**Business Rule from Data mining:**

| **Customer Condition** | **Churn** |
| --- | --- |
| Month-to-month, **high** MonthlyCharges, **low** tenure | Churn |
| Month-to-month, **low** MonthlyCharges | No Churn |
| One-year or Two-year contract | No Churn |

**Conclusion**: Customers on **monthly contracts**, paying **high monthly fees**, and **new to the service (low tenure)** → **high** likelihood of churn.

***END***