ĐẠI HỌC QUỐC GIA TPHCM – TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN – BỘ MÔN HỆ THỐNG THÔNG TIN

A picture containing logo

Description automatically generated

**BÀI TẬP DATA MINING**

**Môn học: Hệ thống thông tin phục vụ trí tuệ kinh doanh**

*GVHD: Hồ Thị Hoàng Vy*

|  |  |  |
| --- | --- | --- |
| **Mã nhóm** | **MSSV** | **Họ và tên** |
| **TTKD-24** | 18120227 | Phạm Văn Minh Phương |
| 18120299 | Trương Công Quốc Cường |
| 18120035 | Đoàn Nguyễn Tấn Hưng |
| 18120534 | Hoàng Công Sơn |

MỤC LỤC

[1 Chuẩn bị 3](#_Toc90916737)

[2 Tạo Data Source 3](#_Toc90916738)

[3 Tạo Data Source View 6](#_Toc90916739)

[4 Đặt yêu cầu và phân tích 8](#_Toc90916740)

[5 Tạo Mining Structure 9](#_Toc90916741)

[6 Đánh giá model 16](#_Toc90916742)

[6.1 Decision Tree 16](#_Toc90916743)

[6.2 Dependency Network 17](#_Toc90916744)

[6.3 Mining Accuracy Chart 18](#_Toc90916745)

[6.3.1 Chọn Input 18](#_Toc90916746)

[6.3.2 Classification Matrix 19](#_Toc90916747)

[6.3.3 Lift Chart 20](#_Toc90916748)

[6.4 Mining Model Prediction 20](#_Toc90916749)

[6.4.1 Khởi tạo 20](#_Toc90916750)

[6.4.2 Kết quả 23](#_Toc90916751)

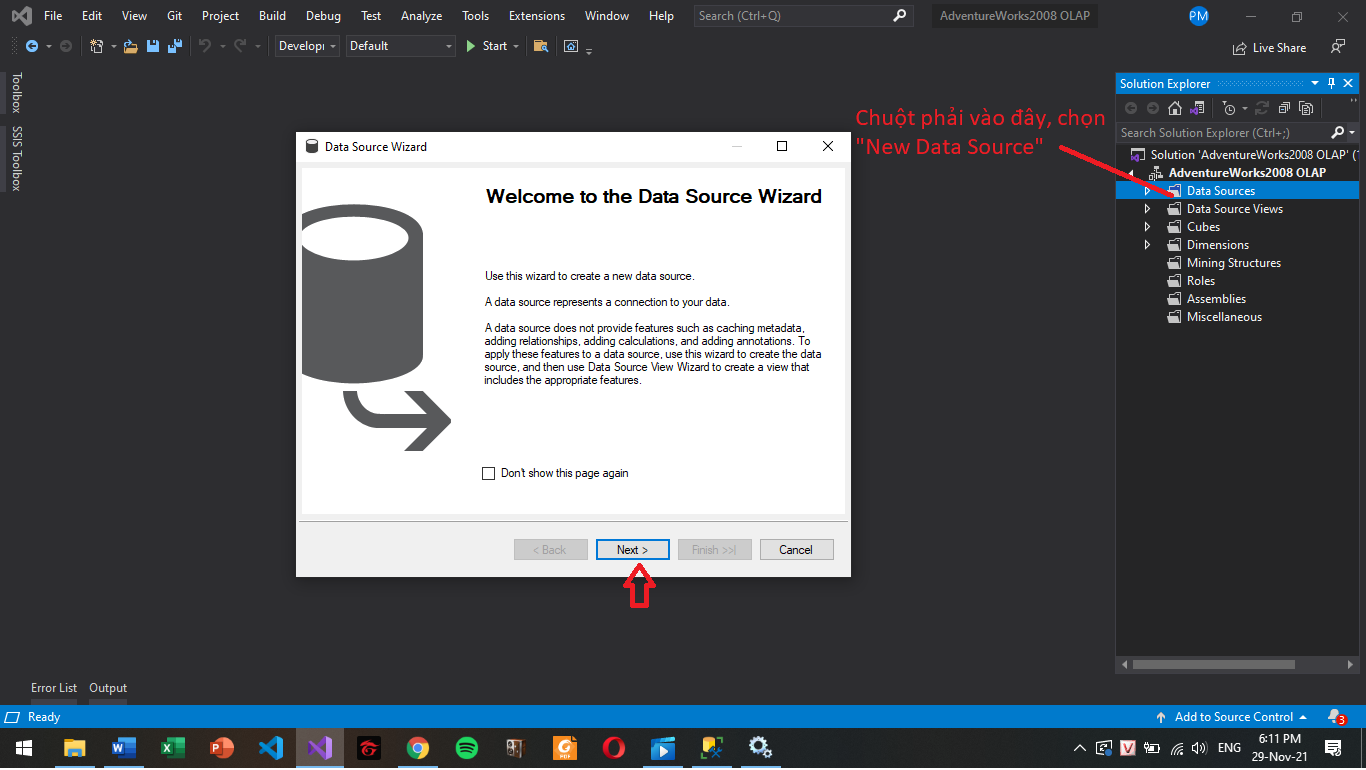
[7 Tham khảo 24](#_Toc90916752)

# Chuẩn bị

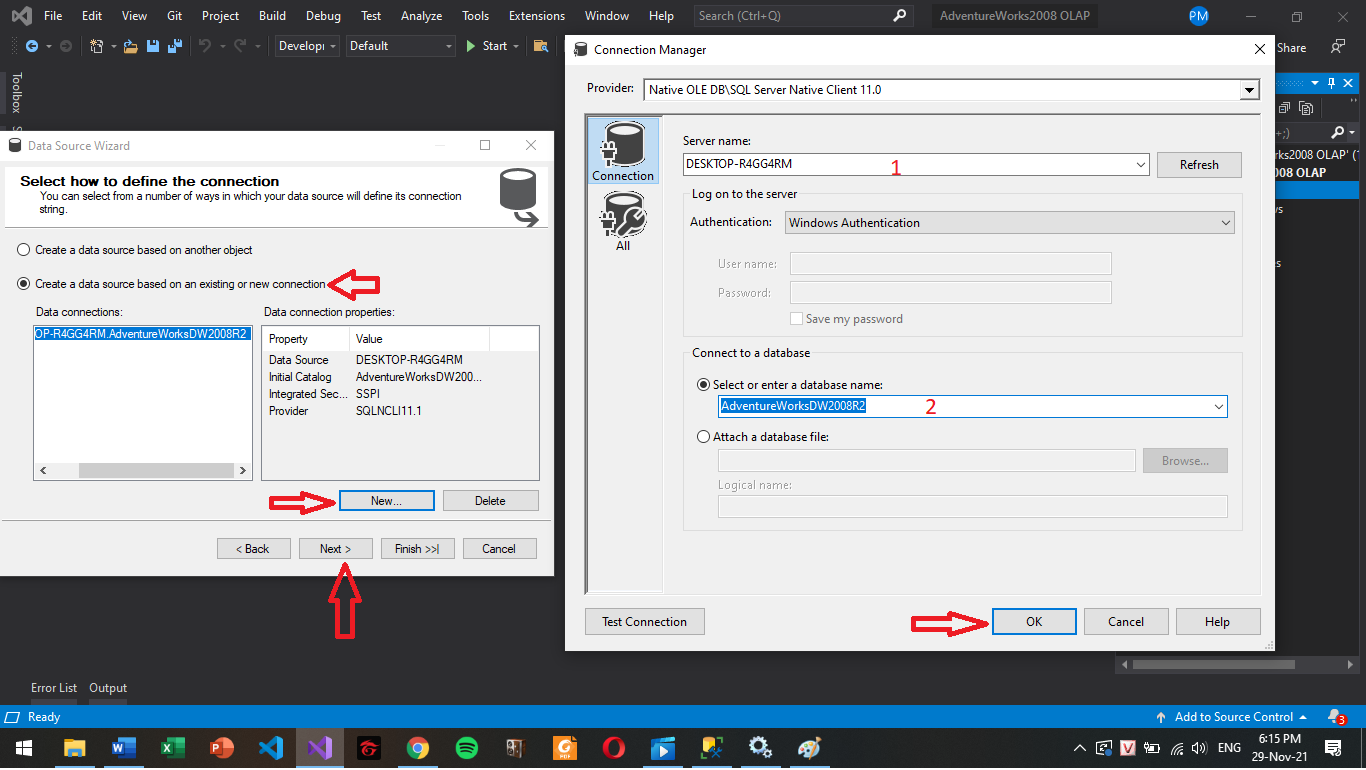
* Cài đặt Analysis Service cho SQL Server
* Nạp database AdventureWorks 2008R2 vào SQL Server
* Chuẩn bị các định danh có quyền truy cập vào các service, database liên quan
* Cài đặt Intergration Service cho Visual Studio

# Tạo Data Source

Tạo datasource cho project bằng cách kết nối tới dữ liệu AdventureWorks mẫu đã nạp sẵn vào SQL Server từ trước.



Tạo kết nối tới dữ liệu có sẵn trong SQL Server và nhấn Next.



QUAN TRỌNG: Sử dụng một định danh có quyền truy cập vào Analysis Service để kết nối vào datasource. Nhấn next.

A screenshot of a computer

Description automatically generated

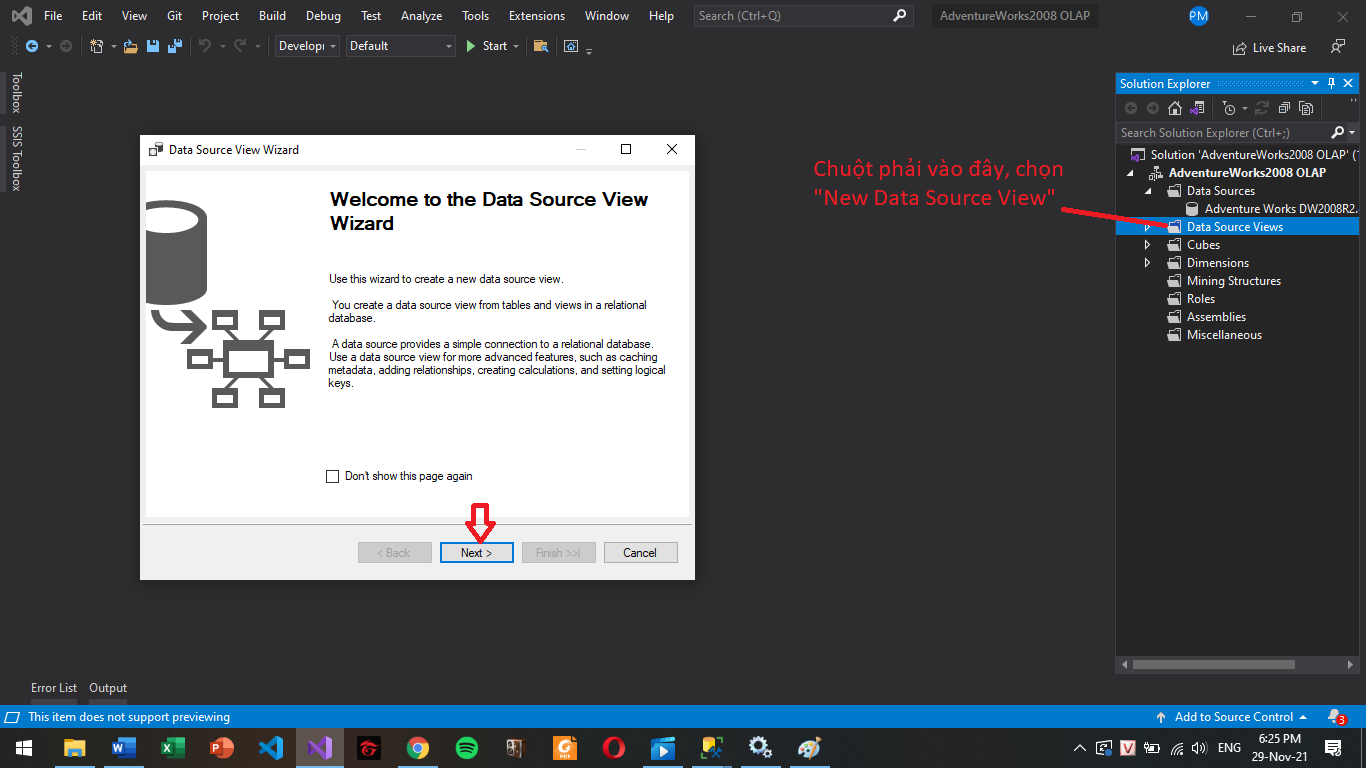
Đặt tên Datasource và nhấn Finish.

A screenshot of a computer

Description automatically generated with medium confidence

# Tạo Data Source View

Tạo data source view



Chọn data source đã tạo ở mục 1. Nhấn next.

A screenshot of a computer

Description automatically generated

Thêm các bảng fact và dim cần thiết vào trong view. Nhấn next

Graphical user interface, application

Description automatically generated

Đặt tên cho view. Nhấn Finish.

A screenshot of a computer

Description automatically generated

# Đặt yêu cầu và phân tích

Yêu cầu: Công ty Adventure Works là một công ty chuyên bán xe đạp và các đồ dã ngoại cần sử dụng dữ liệu để xác định các khách hàng tiềm năng để tiến hành gửi email giới thiệu sản phẩm xe đạp mới tới họ.

Dữ liệu hiện có: Lịch sử mua hàng và các thông tin cá nhân của từng khách hàng gồm tên, tuổi, địa chỉ, thu nhập bình quân, học thức, số con cái, số phương tiện sở hữu…

Mục tiêu: Tìm hiểu xem các thông tin trên có thể có ảnh hưởng gì tới quyết định mua hàng của khách hàng hay không để từ đó chọn ra các khách hàng tiềm năng nhất.

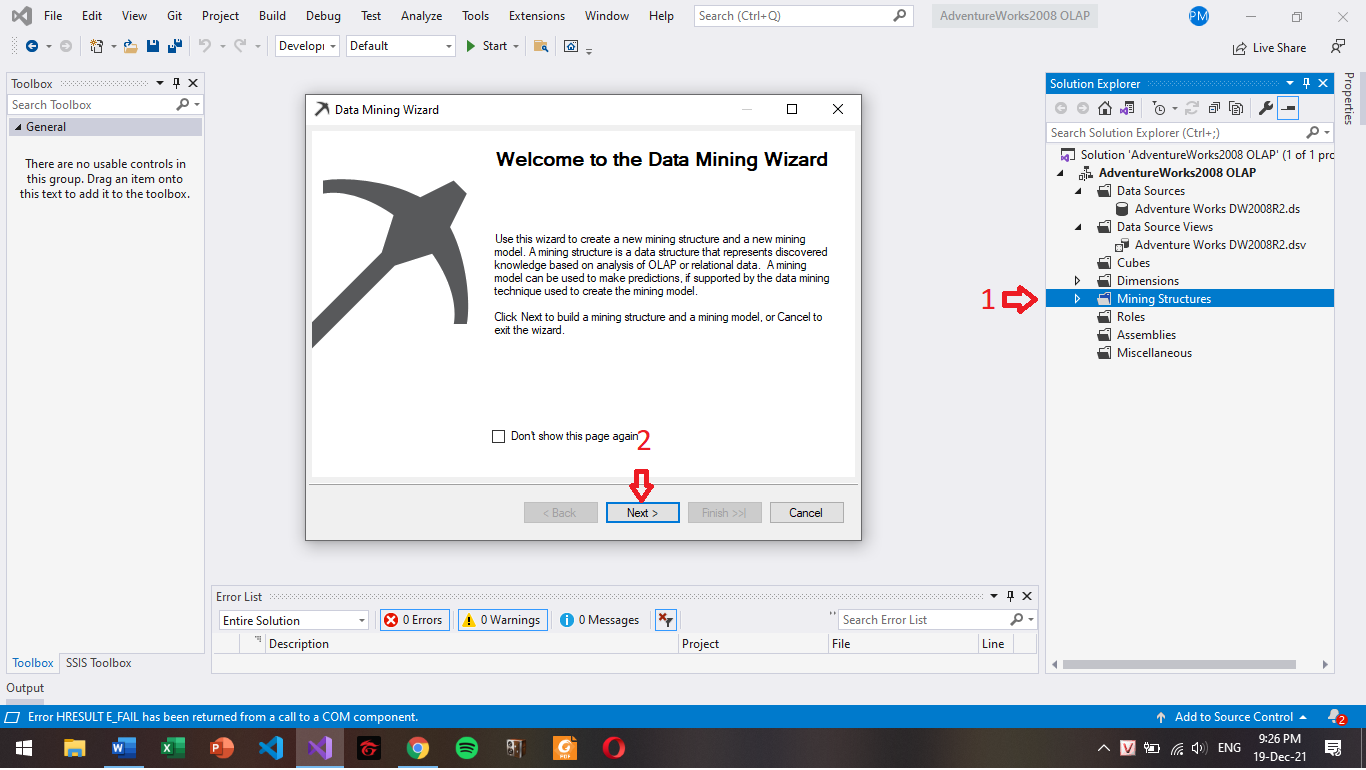
🡪 Phân biệt khách hàng tiềm năng dựa theo các thuộc tính và quy luật. Model mining phù hợp có thể chọn là Decision Tree.

Timeline

Description automatically generated

# Tạo Mining Structure

Bước 1: Chuột phải vào Mining Structure, chọn “New Mining Structure”



Bước 2: Chọn sử dụng data warehouse hoặc CSDL quan hệ có sẵn

Graphical user interface, text, application

Description automatically generated

Bước 3: Chọn cấu trúc mining

Graphical user interface, text, application

Description automatically generated

Bước 4: Chọn Data Source View đã tạo ở mục [[3]](#_Tạo_Data_Source)

Graphical user interface, text, application

Description automatically generated

Bước 5: Chọn view vTargetMail (View xem dữ liệu các khách hàng từng mua xe đạp tại cửa hàng) để làm dữ liệu phân tích

Graphical user interface, application

Description automatically generated

Bước 6: Chọn khóa của bảng, các thông tin có thể ảnh hưởng tới quyết định mua xe (tích ô “Input”) (Ở đây chọn: Tuổi, Thu nhập hàng năm, Số xe hơi sở hữu, Giới tính, Số con cái ở nhà…), và thông tin cần dự đoán (Theo yêu cầu đặt ra là người mua xe – Bike Buyer, tích ô “Predict”)

Graphical user interface, text, application

Description automatically generated

Cấu trúc view vTargetMail:

Graphical user interface, application

Description automatically generated

Bước 7: Xác định kiểu tính chất của dữ liệu, trong trường hợp này đa phần là Rời rạc (Discrete)

Graphical user interface, application

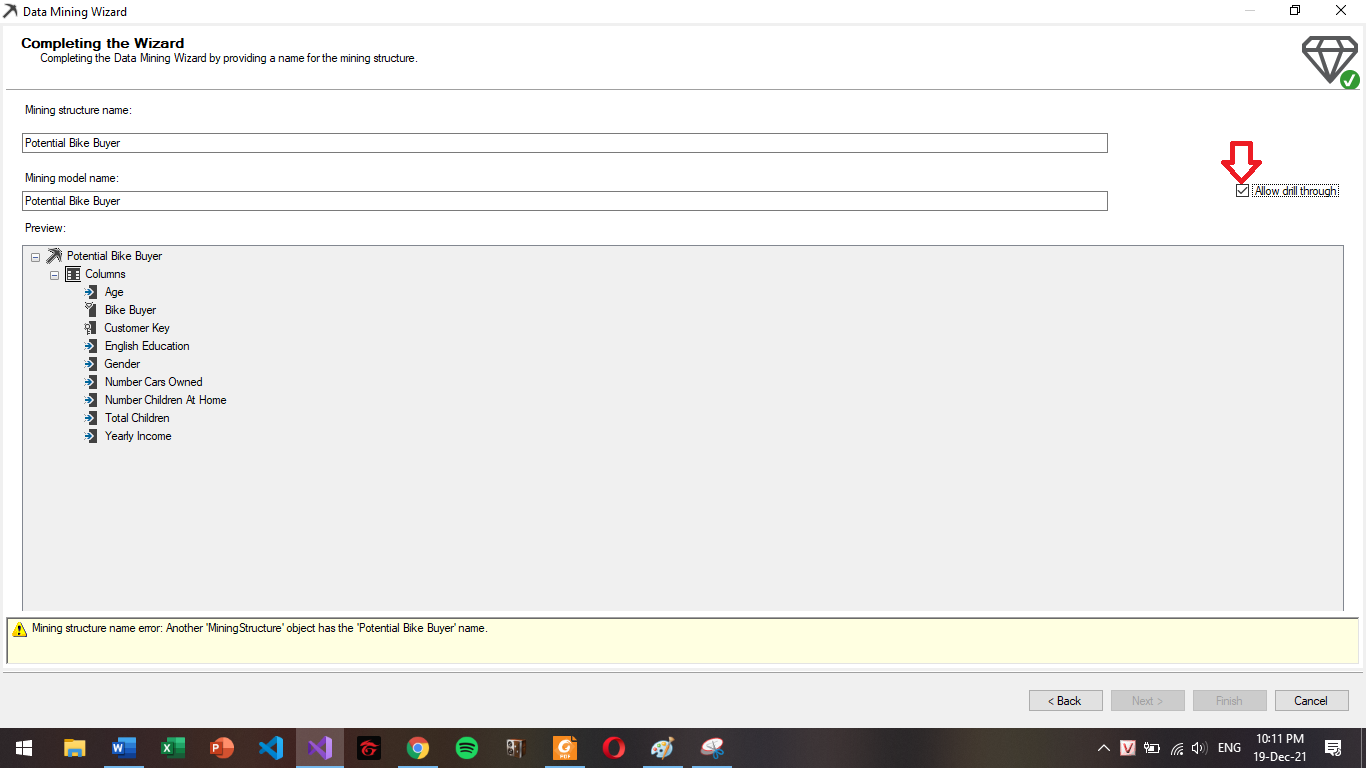
Description automatically generated

Bước 8: Xác định tỉ lệ dữ liệu dùng để đánh giá độ chính xác của model. Ở đây là 30% (Tức 70% dữ liệu dùng để dự đoán, và 30% dữ liệu còn lại được dùng để kiểm tra kết quả)

Graphical user interface, text, application, Word

Description automatically generated

Bước 9: Đặt tên cho mining structure và mining model. Tích chọn “Allow drill through” để cho phép tìm dữ liệu ở cấp chi tiết hơn (nếu có). Và nhấn Finish.



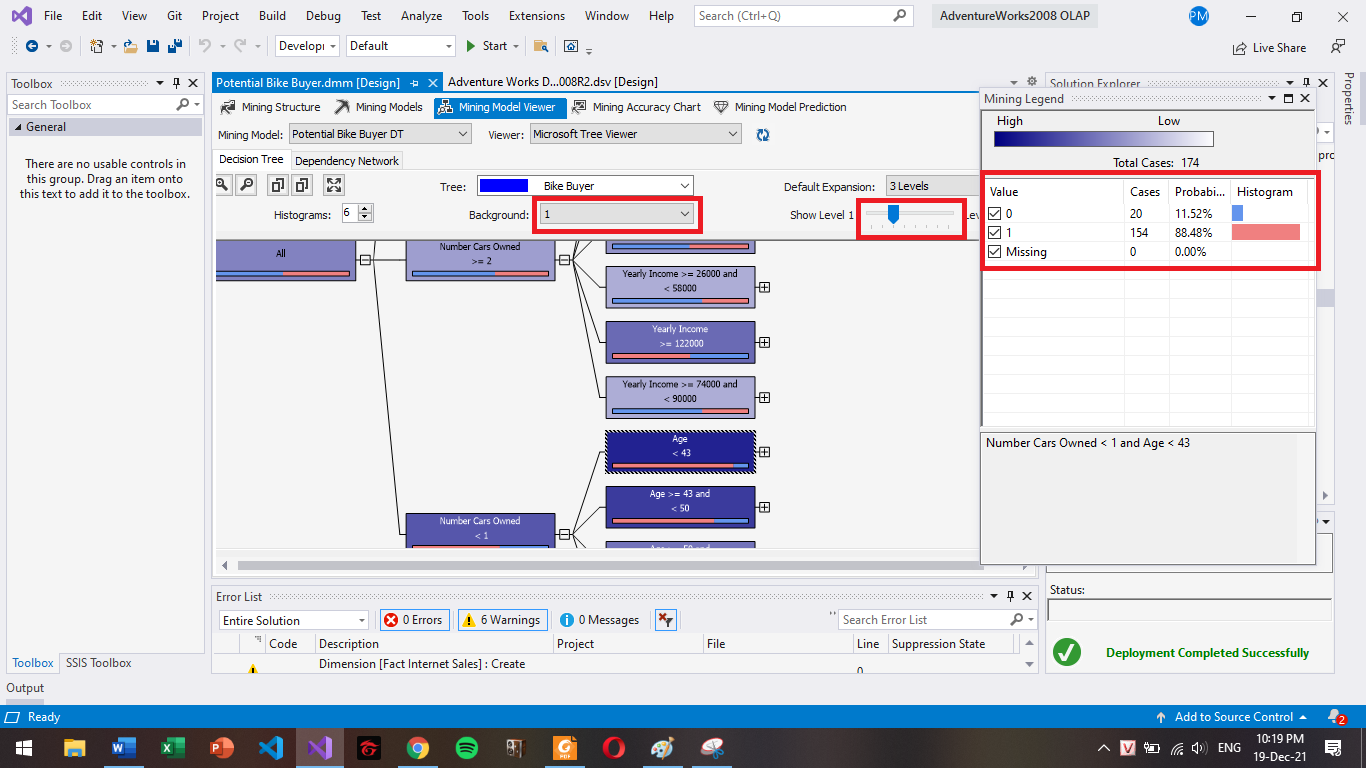
Bước 10: Khởi chạy model

# Đánh giá model

## Decision Tree

Dựa theo các thuộc tính đã chọn như trên, ta có một cây quyết định 8 bậc. Ta có thể chọn background là 1 (Tức thuộc tính Bike Buyer có giá trị 1 – người đã mua xe đạp), show level là 3. Trên sơ đồ, các hình khối có màu càng đậm càng biểu thị tỉ lệ mua xe cao trong nhóm người đó.

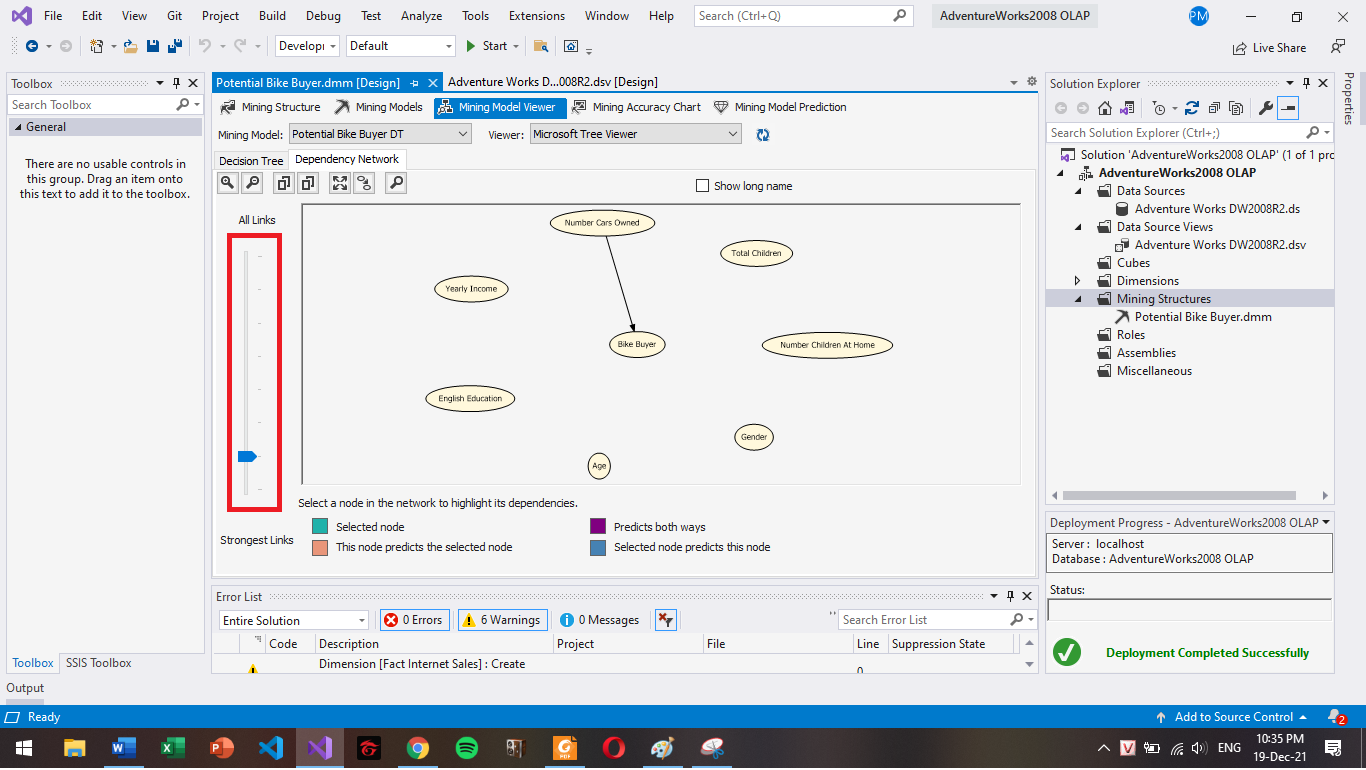
Như trong hình dưới đây, nhóm người sở hữu ít hơn 1 xe hơi và có tuổi < 43 có tỉ lệ mua hàng rất cao, tới 88.48%. Ta có thể mở rộng các phân cấp để xác định cụ thể hơn.

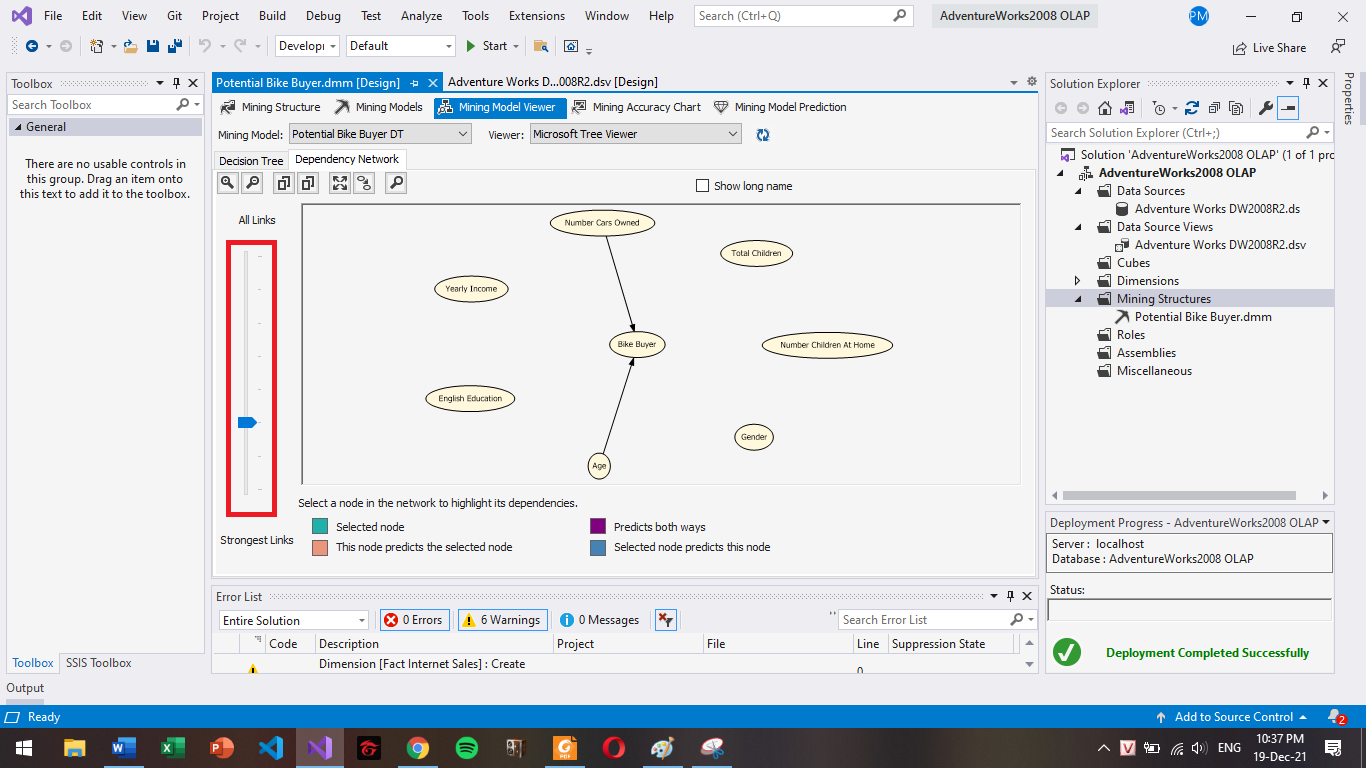


## Dependency Network

Sử dụng dependency network được tạo ra bởi model, ta có thể xác định xem yếu tố nào có ảnh hưởng mạnh hơn tới quyết định mua xe.

Kéo thanh hiển thị các phụ thuộc trong hình, ta có thể xem model cho rằng yếu tố nào sẽ ảnh hưởng mạnh hơn tới quyết định mua xe đạp của người dùng





Theo như model ta có mức độ phụ thuộc (ảnh hưởng đến quyết định mua xe đạp của khách hàng) xếp hạng như sau:

Số xe hơi sở hữu > Tuổi > Thu nhập hàng năm > Số con ở nhà > Tổng số con > Trình độ học vấn > Giới tính.

Điều này khá hợp lý vì ta có thể xét tới một số ảnh hưởng của các yếu tố trên như sau:

- Khách hàng không có xe hơi sẽ mua xe đạp để di chuyển

- Khách hàng trẻ sẽ thích sử dụng xe đạp hơn là các khách hàng lớn tuổi

- Người thu nhấp thấp sẽ dễ chi trả cho xe đạp hơn là các phương tiện khác

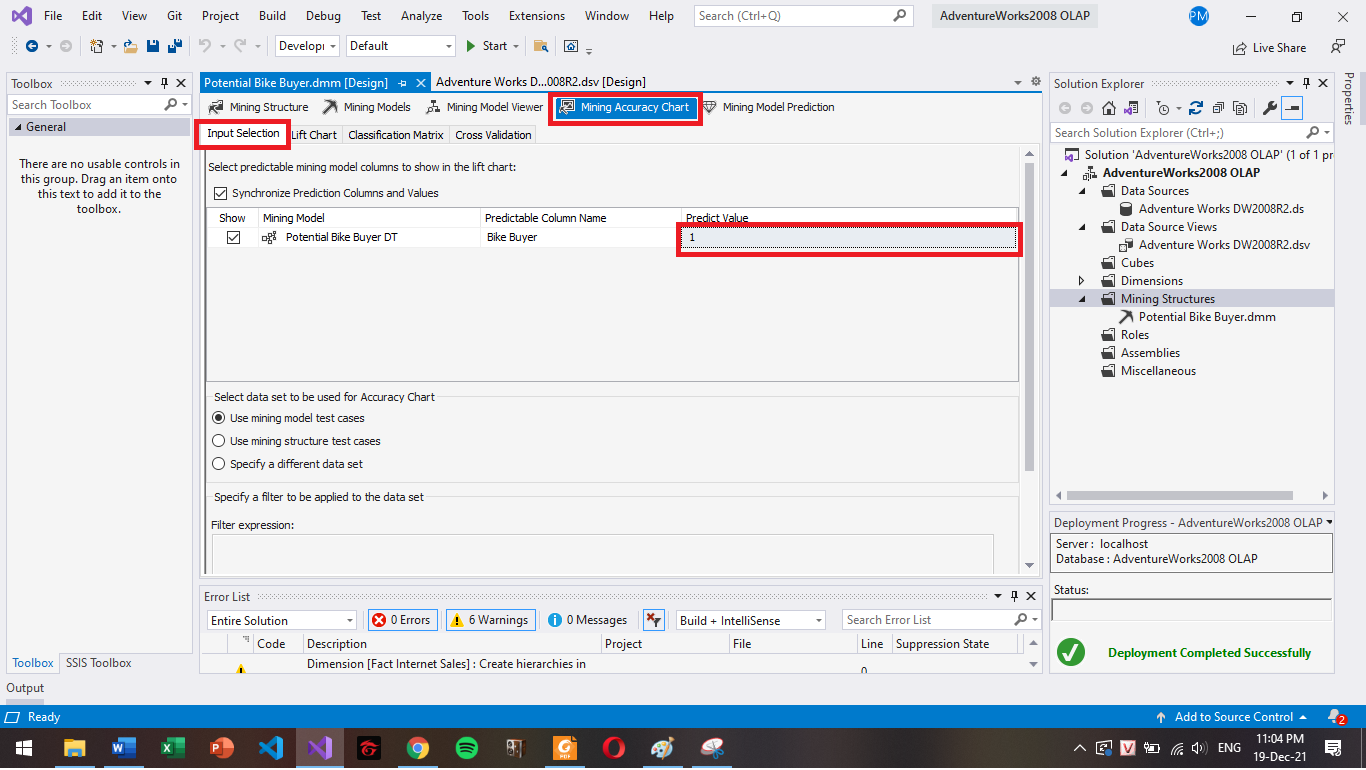
- Khách hàng mua xe đạp cho con

- Trình độ học vấn và giới tính dường như không ảnh hưởng nhiều đến việc họ có muốn một chiếc xe đạp hay không

## Mining Accuracy Chart

### Chọn Input

Chọn Predicttable Column Name là “Bike Buyer” và Predict Value là 1 để dự đoán thuộc tính Bike Buyer



### Classification Matrix

Ta có ma trận phân loại với kết quả như sau:

True Positive Count = 1976

False Positive Count = 910

True Negative Count = 1911

False Negative Count = 748

Graphical user interface, application, Word

Description automatically generated

Accuracy = 

True Positive Rate = 

True Negative Rate = 

Precision = 

Recall = 

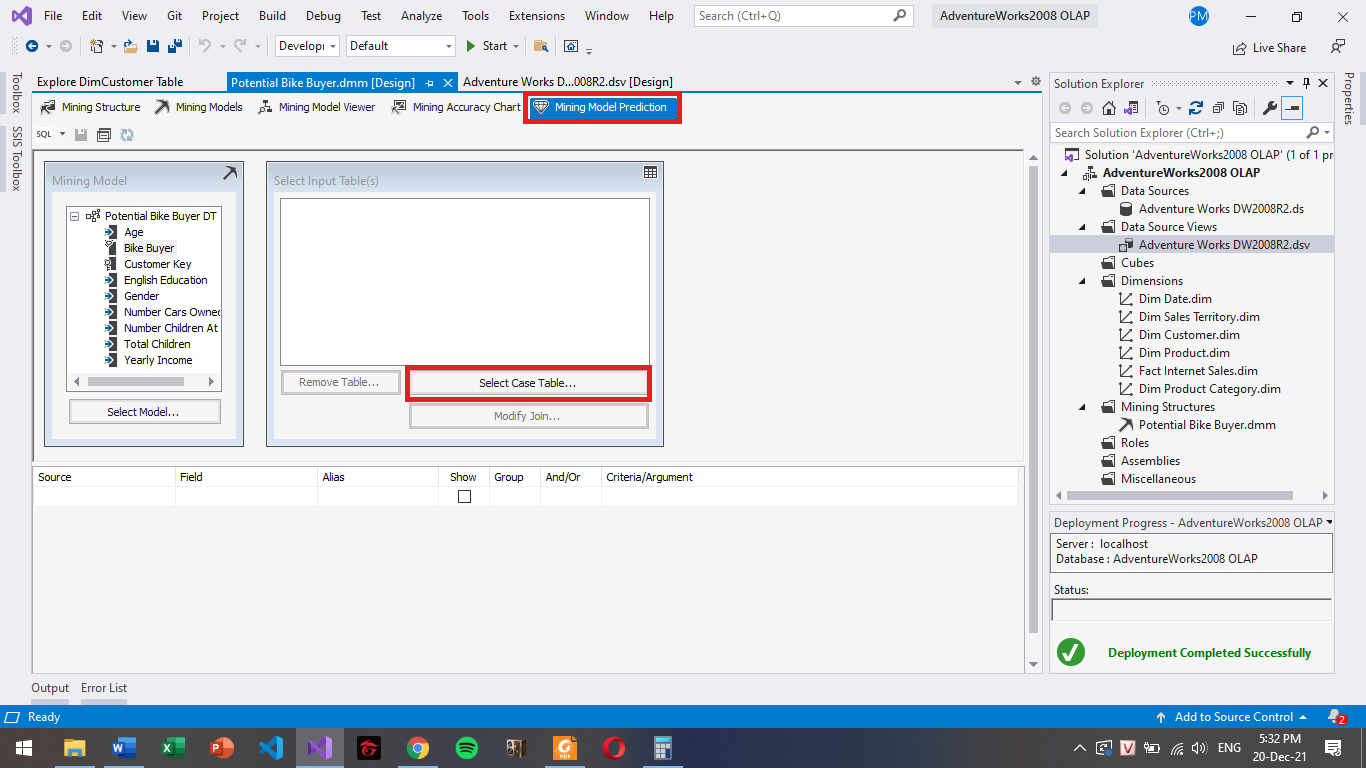
### Lift Chart

Tham khảo mục [[7]](#_Tham_khảo)

## Mining Model Prediction

### Khởi tạo

Bước 1: Mở tab “Mining Model Prediction”, chọn “Select Case Table” để chọn bảng dữ liệu cần dự đoán

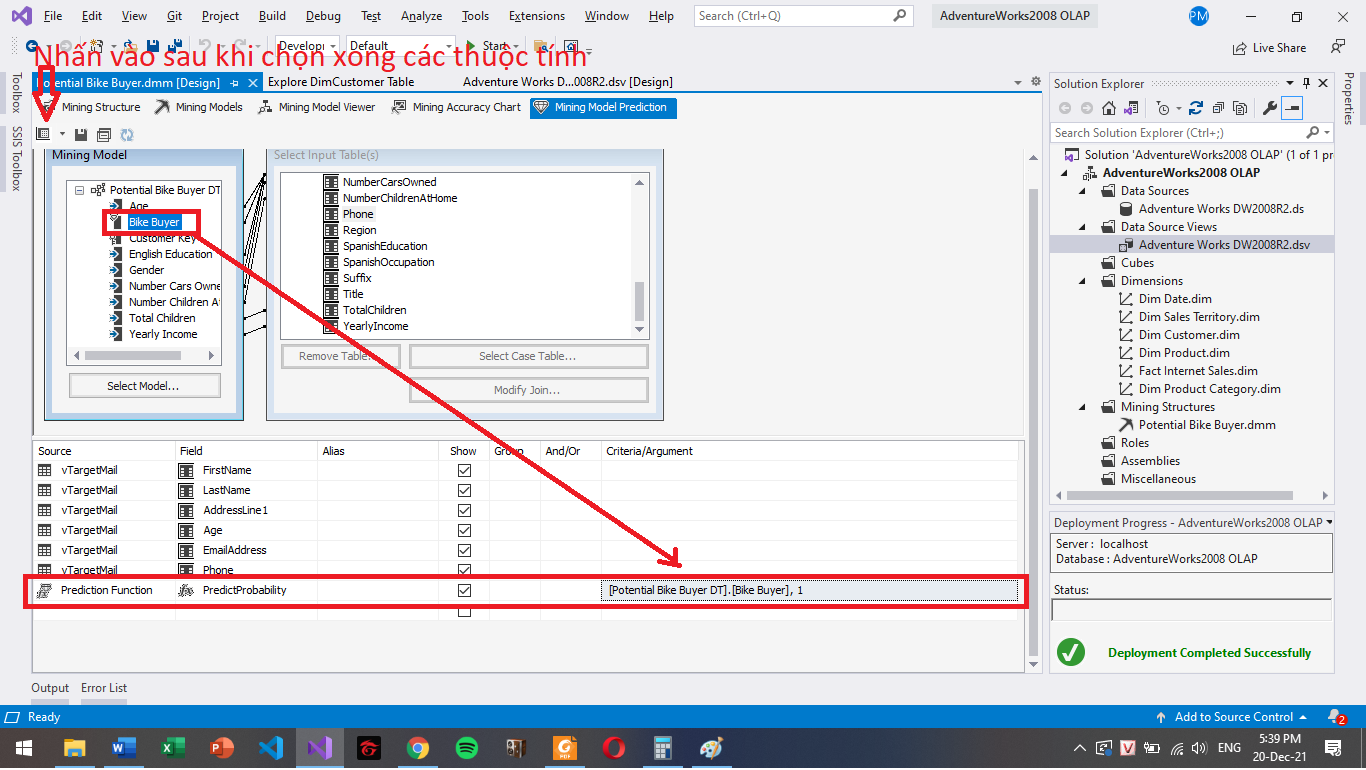


Bước 2: Theo yêu cầu là dự đoán trong danh sách gửi mail khách hàng sẽ mua xe đạp hay không nên ta sẽ chọn bảng khách hàng nào đó. Ở đây vì là dữ liệu backup không có dữ liệu mới nên ta sử dụng lại bảng đã dùng để train model mining là vTargetMail.

Graphical user interface, text, application

Description automatically generated

Bước 3: Double click vào các thuộc tính cần xem để chọn chúng vào câu truy vấn và chọn thêm một Predict Function, chọn “Predict Probability”, kéo thuộc tính Bike Buyer vào Criteria/Argument, viết thêm dấu “,” và tham số 1 để dự đoán tỉ lệ thuộc tính Bike Buyer xuất hiện ra giá trị 1. Tức là tỉ lệ khách hàng cụ thể sẽ mua xe đạp. Nhấn nút “Query” để chạy truy vấn và dự đoán.



### Kết quả

Ta có kết quả dự đoán tỉ lệ một khách hàng sẽ mua xe đạp hay không ở cột Expression

Graphical user interface, application

Description automatically generated

Khi có dữ liệu khách hàng mới thì ta thực hiện quy trình tương tự để xem liệu họ có khả năng sẽ mua hàng hay không

# Tham khảo

<https://www.youtube.com/watch?v=S_j3tDFnwO8>

<https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=2498&context=iemssconference>

https://docs.microsoft.com/en-us/analysis-services/data-mining/lift-chart-analysis-services-data-mining?view=asallproducts-allversions