TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BÀI TẬP LỚN GIỮA KÌ MÔN NHẬP MÔN HỌC MÁY

Người hướng dẫn: **GV.LÊ ANH CƯỜNG**

Người thực hiện: **NGUYỄN THANH TÚ – 52100349**

TRẦN THỊ VẸN –52100674

Lớp **: 21050301**

Khoá **: 25**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM
TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG
KHOA CÔNG NGHỆ THÔNG TIN

BÀI TẬP LỚN GIỮA KÌ MÔN NHẬP MÔN HỌC MÁY

Người hướng dẫn: **GV.LÊ ANH CƯỜNG**

Người thực hiện: **NGUYỄN THANH TÚ – 52100349**

TRẦN THỊ VẸN –52100674

Lớp **: 21050301**

Khoá **: 25**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023

LỜI CẢM ƠN

Tôi xin gửi lời cảm ơn sâu sắc đến Thầy về việc hướng dẫn và giáo dục trong suốt môn học Máy học trong học kỳ vừa qua. Môn học không chỉ giúp tôi nắm vững các khái niệm cơ bản của Máy học mà còn mở ra cho tôi một thế giới mới về ứng dụng của nó trong thực tế.

Thầy đã tạo ra một môi trường học tập tích cực và truyền cảm hứng cho tôi để tiếp cận với những khía cạnh phức tạp và thú vị của Máy học. Những bài giảng thú vị cùng với ví dụ thực tế đã giúp tôi áp dụng kiến thức một cách hiệu quả trong thực tế.

Tôi cũng rất biết ơn về sự tận tâm và tâm huyết của Thầy trong việc hỗ trợ và giải đáp mọi thắc mắc của chúng tôi. Thầy đã không ngừng khích lệ và động viên chúng tôi để chúng tôi có thể vượt qua những thách thức và phát triển tốt hơn trong môn học này.

Môn học không chỉ giúp tôi nắm vững kiến thức mà còn giúp tôi phát triển kỹ năng phân tích và giải quyết vấn đề một cách chủ động. Tôi hy vọng có cơ hội tiếp tục học hỏi từ Thầy và áp dụng kiến thức đã học được vào thực tế công việc sau này.

Một lần nữa, xin chân thành cảm ơn Thầy về sự hỗ trợ và sự dạy dỗ tận tâm của Thầy.

Trân trọng,

ĐỒ ÁN ĐƯỢC HOÀN THÀNH TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi / chúng tôi và được sự hướng dẫn của GV. Lê Anh Cường;. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày 22 tháng 10 năm 2023

Tác giả

(ký tên và ghi rõ họ tên)

Nguyễn Thanh Tú

Trần Thị Vẹn

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

Phần xác nhận của GV hướng dẫn

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

Phần đánh giá của GV chấm bài

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

TÓM TẮT

Câu hỏi 3 : Tìm hiểu vấn đề Overfitting và các phương pháp giải quyết vấn đề này. Yêu cầu bao gồm mô tả lý thuyết và minh hoạ phương pháp trong lý thuyết bằng file code (.ipynb).

MỤC LỤC

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC KÝ HIỆU

CÁC CHỮ VIẾT TẮT

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ

DANH MỤC HÌNH

DANH MỤC BẢNG

CÂU 2 – GIẢI QUYẾT BÀI TOÁN BẰNG TIẾP CẬN HỌC MÁY

2.1 Mô tả bài toán

Bài toán customer churn (hoặc đổi mạng) là một trong những bài toán quan trọng trong lĩnh vực phân tích dữ liệu và quản lý khách hàng. Nó liên quan đến việc dự đoán khả năng một khách hàng hoặc người dùng sẽ chuyển đổi sang dịch vụ hoặc sản phẩm của đối thủ hoặc dừng sử dụng dịch vụ của bạn. Đây là một vấn đề quan trọng đối với các công ty do ảnh hưởng lớn đến doanh thu và lợi nhuận.

Bài toán customer churn thường được tiếp cận thông qua việc phân tích dữ liệu lịch sử của khách hàng, bao gồm thông tin về hành vi sử dụng dịch vụ, thông tin cá nhân, mô hình sử dụng sản phẩm, và bất kỳ dữ liệu nào liên quan khác. Mục tiêu của bài toán là xây dựng mô hình dự đoán khả năng churn của mỗi khách hàng dựa trên dữ liệu hiện có, từ đó giúp doanh nghiệp có thể thực hiện các biện pháp cụ thể để giảm tỷ lệ churn.

Các phương pháp thường được sử dụng để giải quyết bài toán customer churn thuộc dạng phân loại bao gồm sử dụng các thuật toán học máy như logistic regression, decision trees, và các phương pháp kết hợp khác. Quá trình này thường bao gồm các bước như tiền xử lý dữ liệu, chia tập dữ liệu thành tập huấn luyện và tập kiểm tra, xây dựng mô hình, đánh giá và tinh chỉnh mô hình để đạt được hiệu suất tốt nhất.

Việc giải quyết bài toán customer churn có thể giúp các doanh nghiệp phát hiện ra các yếu tố ảnh hưởng đến quyết định của khách hàng và đưa ra các chiến lược hợp lý để giữ chân khách hàng hiện tại, từ đó nâng cao hiệu quả kinh doanh và tăng trưởng doanh số.

2.2 Các feature trong bài toán

Bài toán customer churn thường có đa dạng loại feature, bao gồm cả các loại feature dạng số (numerical) và các loại feature dạng phân loại (categorical). Đa dạng loại feature này giúp tái hiện rõ hơn hành vi của khách hàng và từ đó tạo ra các mô hình dự đoán churn chính xác hơn.

2.2.1 Categorical Features (Đặc trưng Phân loại):

- Gender
- SeniorCitizen
- Partner
- Dependents
- PhoneService
- MultipleLines
- InternetService
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV
- StreamingMovies
- Contract
- PaperlessBilling
- PaymentMethod

2.2.3 Numerical Features (Đặc trưng Số):

- Tenure (thời gian sử dụng dịch vụ tính bằng tháng)
- MonthlyCharges (cước hàng tháng)
- TotalCharges (tổng cộng cước phí)

Kết hợp các loại feature này trong mô hình học máy giúp mô hình có thể hiểu rõ hơn về hành vi và đặc điểm của từng khách hàng, từ đó tạo ra các dự đoán chính xác hơn về khả năng churn của họ. Qua đó, doanh nghiệp có thể áp dụng các biện pháp cụ thể để duy trì và tăng cường mối quan hệ với khách hàng.

2.3 Feature selection

Khi sử dụng SelectKBest để lựa chọn ra các đặc trưng quan trọng, chúng ta cần xác định trước số lượng đặc trưng muốn chọn. Dưới đây là cách bạn có thể phân loại feature đã cung cấp sử dụng SelectKBest với số lượng đặc trưng 16 được chọn:

2.3.1 Categorical Features (Đặc trưng Phân loại):

- SeniorCitizen
- Partner
- Dependents
- InternetService
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV
- Contract
- PaperlessBilling
- PaymentMethod

2.3.2 Numerical Features (Đặc trưng Số):

- Tenure (thời gian sử dụng dịch vụ tính bằng tháng)
- MonthlyCharges (cước hàng tháng)
- TotalCharges

Dựa trên kết quả đánh giá các mô hình, có một số sự cải thiện nhất định trong hiệu suất của các mô hình sau khi áp dụng SelectKBest với phương pháp f_classif. Tuy nhiên, mức độ cải thiện này có thể khá khác nhau đối với từng mô hình cụ thể. Chẳng hạn, mô hình Logistic Regression đã cho thấy sự cải thiện đáng kể trong các độ đo đánh giá, trong khi các mô hình khác có sự cải thiện nhỏ hơn hoặc không đáng kể.

Trong quá trình feature selection, mục tiêu chính là loại bỏ những đặc trưng không quan trọng hoặc không cần thiết để giảm chiều của dữ liệu mà vẫn duy trì hoặc cải

thiện hiệu suất của mô hình. Tuy nhiên, việc lựa chọn đặc trưng không phải lúc nào cũng dẫn đến cải thiện hiệu suất của mô hình, đặc biệt là khi dữ liệu ban đầu đã được chuẩn bị kỹ lưỡng trước khi áp dụng mô hình.

Do đó, việc sử dụng phương pháp feature selection như SelectKBest có thể cải thiện hiệu suất của mô hình trong một số trường hợp, đặc biệt là khi dữ liệu có quá nhiều đặc trưng không cần thiết hoặc không quan trọng. Tuy nhiên, việc lựa chọn phương pháp phù hợp và kiểm soát kỹ thuật feature selection là cần thiết để đảm bảo rằng việc giảm chiều dữ liệu không làm mất mát thông tin quan trọng và không ảnh hưởng đến hiệu suất của mô hình.

TÀI LIỆU THAM KHẢO

Tiếng Việt

Tiếng Anh

PHỤ LỤC