BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP THÀNH PHỐ HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

ĐỀ TÀI: ỨNG DỤNG CÂY QUYẾT ĐỊNH ĐỂ PHÂN LOẠI NGƯỜI DÙNG TRONG THƯƠNG MẠI ĐIỆN TỬ

Giáo viên hướng dẫn : Nguyễn Thị Mỹ Linh

Họ và tên sinh viên : Nguyễn Vũ Khánh Huy MSSV: 16025591
Lớp: DHKHMT12A

1.Mô tả đề tài

Giải quyết các vấn đề trong E-tailing mà cụ thể là bài toán phân loại người dùng. Dựa vào cây quyết định, để tìm ra sự liên quan giữa các thuộc tính trên một hóa đơn và hành vi tiêu dùng của người dùng trên sàn thương mại điện tử

2. Lý do chọn đề tài

Phân loại người dùng là giải pháp tăng doanh thu cho các doanh nghiệp khi giúp họ hiểu và phục vụ chính xác hơn cho nhu cầu của từng người dùng.

3. Mục tiêu đề tài

Áp dụng cây quyết định để phân loại người dùng cùng các thuật toán được cải tiến từ cây quyết định như random forest, adaboost và các thuật toán không liên quan đến cây quyết định như logistic regression, KNN để áp dụng và so sánh trên tập người dùng 500000 dòng.

4.Phương pháp nghiên cứu

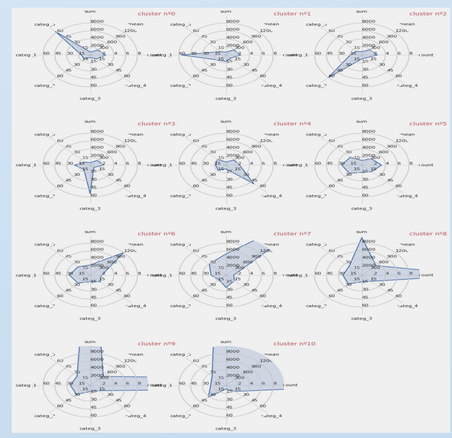
Phân tích để tìm thấy sự liên quan giữa các thuộc tính trong tập dữ liệu, sử dụng kmeans để phân cụm sản phẩm, và cụm người dùng. Phân loại người dùng dựa trên cụm sản phẩm và cụm người dùng đã có được.

5. Kế hoạch thực hiện

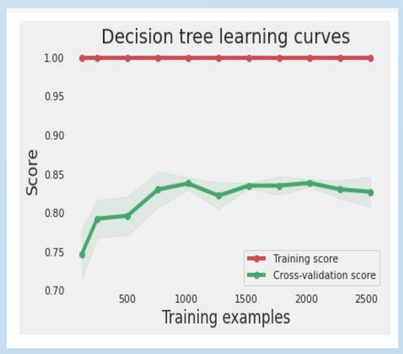
1.Chuẩn bị dữ liệu.
2.Tìm hiểu về nội dung của tập dữ liệu.
3.Tiếp cận phân loại sản phẩm.
4.Tiếp cận phân loại người dùng.
5.Thử phân loại người dùng trên nhiều thuật toán khác nhau.
6.Kết luận.

6. Kết quả

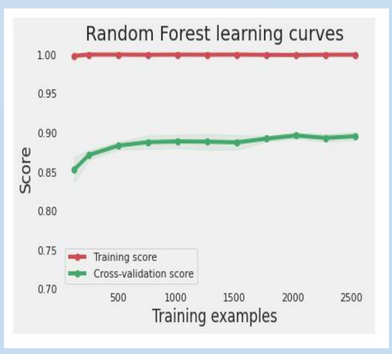
Hình ảnh 5 cụm sản phẩm .



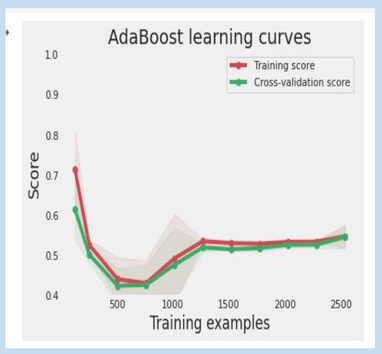
Hình ảnh 11 cụm người dùng.



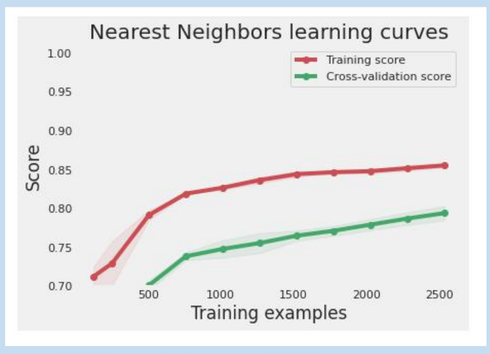
Decision tree cho độ chính xác 86.13%



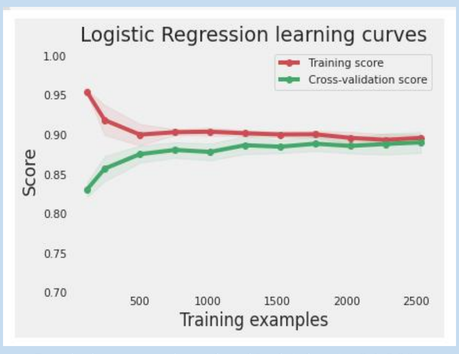
Random forest cho độ chính xác 92.56%



Adaboost cho độ chính xác 51.20%



KNN cho độ chính xác 79.78%.



Logistic regression cho độ chính xác 86.29%.

7. Kết luận: Bằng các kết quả kiểm thử , ta thấy cây quyết định là một thuật toán phụ thuộc rất lớn vào tập dữ liệu. Thậm chí với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc và mô hình cây có thể thay đổi hoàn toàn. Overfitting là vấn đề thường xuyên gặp phải kể cả khi sử dụng các thuật toán tiên tiến hơn như random forest, adaboost cũng vẫn gặp phải.

Tài liệu tham khảo .

Tài liệu Tiếng Việt.
[1]Bách khoa toàn thư mở wikipedia, thương mại điện tử.
[2]ThS. Nguyễn Vương Thịnh, bài giảng khai phá dữ liệu của ĐH Bách khoa TP HCM.
[3]Vũ Hữu Tiệp, decision tree, iterative dichomister 3, machine learning cơ bản.
Tài liệu nước ngoài .
[4]Breiman, L., J. Friedman, R. Olshen, R., and Stone, C. (1984). Classification and regression trees. Wadsworth Books
[5]Teli, S., Kanikar, P. (2015). A survey on decision tree based approaches in data mining.International Journal of Advanced Researches in Computer Science and Software Engineering, Volume 5, Issue 4.
[6] Fabien Daniel, Kaggle E-commerce, actual transaction from UK retailer.
[7] Peter J. Rousseeuw (1987). “Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis”. Computational and Applied Mathematics 20: 53-65.

Email : khanhhuynguyenvu@gmail.com Phone: 0938294687.