Tóm tắt bài báo

1. Tên đồ án (bài báo):  
Machine learning-based e-commerce platform repurchase customer prediction model  
(Mô hình dự đoán khách hàng mua lại cho nền tảng thương mại điện tử dựa trên học máy)

2. Tác giả:  
- Cheng-Ju Liu  
- Tien-Shou Huang  
- Ping-Tsan Ho  
- Jui-Chan Huang  
- Ching-Tang Hsieh

3. Nơi công bố:  
- Tạp chí PLOS ONE (Hoa Kỳ)

4. Năm công bố:  
- 2020 (Ngày xuất bản: 3/12/2020)

5. Mục tiêu nghiên cứu:  
- Phân tích những hạn chế của các phương pháp dự đoán hành vi mua hàng trực tuyến truyền thống.  
- Đề xuất một hệ thống phân tích và dự đoán hành vi mua lại của khách hàng trên nền tảng thương mại điện tử.  
- So sánh hiệu quả giữa các mô hình học máy và tối ưu hóa dự đoán bằng kỹ thuật kết hợp mô hình.

6. Cơ sở lý thuyết:  
- Học máy (Machine Learning): Công cụ phân tích dữ liệu mạnh mẽ ứng dụng trong nhiều lĩnh vực.  
- Các thuật toán: Logistic Regression, XGBoost, C4.5 Decision Tree, SVM, Artificial Neural Network.  
- Các yếu tố tâm lý khách hàng: Sự hài lòng khách hàng, Ý định mua lại, Thái độ khách hàng.

7. Phương pháp thu thập và xử lý thông tin:  
- Nguồn dữ liệu: Hành vi người dùng từ nền tảng TMĐT, chia cửa sổ 7 ngày.  
- Xử lý dữ liệu: Downsampling, Feature selection giảm từ 110 xuống 56 đặc trưng.  
- Công cụ: Java + MySQL, R, Python.  
- Đánh giá: AUC, F1 Score, Precision, Recall, 10-fold cross-validation.

8. Kết quả đạt được:  
- XGBoost hoạt động tốt hơn Logistic Regression.  
- Fusion Model cải thiện nhẹ so với XGBoost đơn lẻ.  
- Feature Selection nâng cao F1 Score đáng kể.  
- P/N sampling làm mô hình XGBoost ổn định hơn.  
- SVM kết hợp dao động ổn định cho dự đoán doanh số tốt nhất.

9. Hạn chế của công trình:  
- Fusion model chỉ cải thiện nhẹ so với XGBoost.  
- Linear combination dễ overfitting.  
- Dữ liệu từ một nền tảng duy nhất, tính khái quát chưa cao.  
- Chưa thử nghiệm các kỹ thuật học sâu tiên tiến như LSTM hay Attention-based models.