Dự đoán doanh số đấu giá

Prediction of Auction Sales

Nguyễn Văn Huy - 21130382

Đỗ Xuân Hậu - 21130349

Khoa Công nghê thông tin, Đại học Nông Lâm TP.HCM

Tóm tắt: Dự đoán doanh số đấu giá tập trung vào việc dự đoán xem doanh thu bán hàng của một phiên đấu giá mới có cao hơn doanh thu bán hàng trung bình của danh mục sản phẩm hay không. Mục tiêu của dự án là xây dựng một mô hình dự đoán để đưa ra dự đoán về doanh thu bán hàng thực tế và so sánh nó với doanh thu bán hàng trung bình. Các thuật toán được sử dụng để xây dựng mô hình dự đoán là : Neural Network, SVM, kNN, Naïve Bayes, RandomForest, Decision Tree, Regression và Clustering. Các bước thực hiện dự án bao gồm thu thập dữ liệu về phiên đấu giá và doanh thu bán hàng, tiền xử lý dữ liệu như xử lý dữ liệu thiếu, feature selection và xử lý dữ liệu không cân bằng. Tiếp theo, chúng ta sẽ xây dựng các mô hình dự đoán và điều chỉnh các tham số của chúng. Quá trình huấn luyện, xác thực và kiểm tra mô hình sẽ được thực hiện để đánh giá hiệu suất của chúng. Kết quả của dự án sẽ là một mô hình dự đoán có khả năng dự đoán xem doanh thu bán hàng của phiên đấu giá mới có cao hơn doanh thu bán hàng trung bình hay không. Kết quả sẽ được đánh giá bằng các độ đo như Accuracy, Precision, Recall và F1-score. Cuối cùng, chúng ta sẽ phân tích và đánh giá kết quả đạt được

Abstract: This project focuses on predicting whether the sales revenue of a new auction session will be higher than the average sales revenue of the product portfolio. The objective of the project is to build a prediction model that forecasts the actual sales revenue and compares it with the average sales revenue. The algorithms used to construct the prediction model include Neural Network, SVM, kNN, Naïve Bayes, RandomForest, Decision Tree, Regression, and Clustering. The project implementation steps involve data collection on auction sessions and sales revenue, data preprocessing such as handling missing data, feature selection, and addressing imbalanced data. Next, we will build prediction models and tune their parameters. The training, validation, and testing process will be conducted to evaluate the performance of these models. The project outcome will be a prediction model capable of determining whether the sales revenue of a new auction session is higher than the average sales revenue. The results will be assessed using metrics such as Accuracy, Precision, Recall, and F1-score. Finally, we will analyze and evaluate the achieved results.

1. Giới thiệu

- Dự đoán doanh số đấu giá sẽ dự đoán xem doanh số đấu giá và xem liệu doanh thu bán hàng trong một phiên đấu giá có cao hơn doanh thu bán hàng trung bình của danh mục sản phẩm hay không. Đây là một vấn đề quan trọng trong lĩnh vực kinh doanh và marketing, nơi mà việc dự đoán và ước tính doanh thu bán hàng có thể giúp doanh nghiệp đưa ra các quyết định chiến lược về giá cả, quảng cáo và quản lý sản phẩm.
- Tầm quan trọng và sự cần thiết của vấn đề nghiên cứu trong dự án này nằm ở việc cung cấp cho doanh nghiệp cái nhìn trước về tiềm năng doanh thu của một phiên đấu giá cụ thể. Bằng cách dự đoán liệu doanh thu bán hàng có cao hơn doanh thu bán hàng trung bình hay không, doanh nghiệp có thể điều chỉnh chiến lược kinh doanh, tối ưu hóa nguồn lực và tăng cường hiệu quả tiếp thị.
- Khám phá giá là một khía cạnh cơ bản của bất kỳ cuộc đấu giá nào. Nó liên quan đến việc xác định giá của một tài sản dựa trên động lực cung và cầu. Phân tích dự đoán có thể nâng cao khả năng khám phá giá bằng cách cung cấp thông tin chi tiết theo thời gian thực về điều kiện thị trường. Ví dụ, trong ngành công nghiệp ô tô, dữ liệu về doanh số bán hàng trước đây, nhu cầu thị trường và các chỉ số kinh tế có thể giúp dự đoán giá trị tương lai của phương tiện. Các nhà điều hành đấu giá có thể sử dụng những dự đoán này để đặt giá khởi điểm phản ánh giá trị thị trường thực sự, cải thiện tỷ suất lợi nhuận cuối cùng và tỷ giá bán thông qua. Ngoài ra, trong thị trường hàng hóa kim loại, phân tích dự đoán có thể phân tích chuỗi cung ứng toàn cầu, các yếu tố địa chính trị và dữ liệu giá cả lịch sử để dự báo giá tương lai của các kim loại như nhôm, đồng và thép. Điều này cho phép người bán tính thời gian đấu giá một cách chiến lược và người mua đưa ra quyết định mua hàng sáng suốt.
- Các phương pháp áp dụng để giải quyết bài toán trong dự án này liên quan chủ yếu đến Machine Learning và các thuật toán dự đoán. Các phương pháp như Neural Network, SVM, kNN, Naïve Bayes, RandomForest, Decision Tree, và các thuật toán Regression và Clustering có thể được áp dụng để xây dựng mô hình dự đoán. Qua quá trình tiền xử lý dữ liệu, huấn luyện và đánh giá mô hình, chúng ta có thể đưa ra dự đoán về doanh thu bán hàng và so sánh nó với doanh thu bán hàng trung bình để đưa ra quyết định kinh doanh.

2. Các công trình liên quan

1. ***Predicting Online Auction Prices from Textual Descriptions (Tác giả: Saikat Basu, Anupam Joshi, và Tim Finin)***

- Một nghiên cứu về sử dụng Neural Network để dự đoán giá đấu giá đã được thực hiện, cho thấy Neural Network có khả năng mô hình hóa mối quan hệ phi tuyến giữa các đặc điểm sản phẩm và giá bán.

- Ưu điểm: Sử dụng thông tin từ mô tả văn bản để dự đoán giá đấu giá là một phương pháp tiềm năng và có thể cung cấp thông tin hữu ích cho các nhà bán hàng.
- Nhược điểm: Công trình này tập trung chủ yếu vào dự đoán giá đấu giá và không đề cập đến việc dự đoán doanh thu bán hàng.

2. ***Predicting Sales and Price Elasticity in Online Auctions with Many Item-specific Fixed Effects" (Tác giả: Patrick Bajari, Ali Hortacsu, John Nekipelov, và Steven Tadelis)***

- Phương pháp: Công trình này sử dụng mô hình Fixed Effects Regression để dự đoán doanh số và đàn hồi giá trong các phiên đấu giá trực tuyến. Mô hình xử lý các yếu tố cố định đặc thù của từng sản phẩm thông qua việc tính toán các hiệu ứng cố định riêng biệt cho từng mặt hàng.
- Ưu điểm: Mô hình Fixed Effects Regression có thể xử lý các yếu tố cố định đặc thù của từng sản phẩm, giúp cải thiện độ chính xác của dự đoán doanh thu và đàn hồi giá.
- Nhược điểm: Phương pháp này có thể yêu cầu nhiều dữ liệu và tính toán phức tạp để ước tính hiệu ứng cố định cho từng sản phẩm.

3. ***Forecasting Sales in Retail Using Bayesian Structural Time Series (Tác giả: Taylor, S.J., Letham, B., và van den Berg, E.)***

- Phương pháp: Công trình này sử dụng mô hình Bayesian Structural Time Series (BSTS) để dự đoán doanh số bán hàng trong ngành bán lẻ. Mô hình BSTS kết hợp các yếu tố thời gian, mùa vụ, và các yếu tố đặc thù của từng sản phẩm để tạo ra dự báo chính xác.
- Ưu điểm: Mô hình BSTS có khả năng xử lý các yếu tố thời gian và mùa vụ, đồng thời tích hợp các yếu tố đặc thù của sản phẩm để tạo ra dự đoán doanh thu chính xác trong ngành bán lẻ.
- Nhược điểm: Mô hình BSTS có thể đòi hỏi sự hiểu biết về thống kê và xử lý dữ liệu phức tạp để triển khai và đào tạo mô hình.

3. Phát biểu bài toán

3.1 Bài toán :

- **Giới thiệu bài toán của project:** Bài toán của dự án này là xây dựng mô hình dự đoán doanh thu bán hàng thực tế có cao hơn doanh thu bán hàng trung bình của danh mục sản phẩm hay không trong mỗi phiên đấu giá mới.

Inputs:

- ID của phiên đấu giá.
- Tên danh mục sản phẩm.
- Tiêu đề sản phẩm.
- Phụ đề sản phẩm.
- Ngày bắt đầu phiên đấu giá.

- Ngày kết thúc phiên đấu giá.
- Thời gian đấu giá.
- Mã loại hình đấu giá.
- Điểm phản hồi tại thời điểm đấu giá.
- Giá khởi điểm.
- Giá mua ngay.
- Cờ giá mua ngay.
- Cờ phí in đậm.
- Cờ phí nổi bật.
- Cờ phí nổi bật theo danh mục.
- Cờ phí trưng bày.
- Cờ phí trưng bày nổi bật.
- Cờ phí trưng bày IPX nổi bật.
- Cờ phí dự trữ.
- Cờ phí làm nổi bật.
- Cờ phí lịch trình.
- Cờ phí viền.
- Số lượng sản phẩm có sẵn trong mỗi phiên đấu giá.
- Doanh thu bán hàng.
- Doanh thu trung bình của danh mục sản phẩm.
- Cờ doanh thu cao hơn trung bình.

Outputs:

- Xác định doanh thu bán hàng thực tế có cao hơn doanh thu trung bình của danh mục sản phẩm hay không (1 hoặc 0).

3.2 Thuật toán .

Neural Network

Neural Network là một mạng lưới các nơ-ron được kết nối với nhau, mô phỏng cách hoạt động của não người.

- **Các bước hiện thực:**
    1. **Chuẩn bị dữ liệu:** Xử lý dữ liệu bị thiếu, chuẩn hóa dữ liệu.
    2. **Xây dựng mô hình:** Xác định số lớp và số nơ-ron trong mỗi lớp, hàm kích hoạt.
    3. **Huấn luyện mô hình:** Sử dụng tập dữ liệu huấn luyện, chọn hàm mất mát và tối ưu hóa.
    4. **Đánh giá mô hình:** Sử dụng tập dữ liệu kiểm tra.

SVM

Support Vector Machine (SVM) là một thuật toán phân loại mạnh mẽ, đặc biệt hiệu quả với dữ liệu nhiều chiều.

- **Các bước hiện thực:**

1. **Chuẩn bị dữ liệu:** Chuẩn hóa dữ liệu.
2. **Xây dựng mô hình:** Chọn kernel thích hợp (linear, polynomial, RBF).
3. **Huấn luyện mô hình:** Sử dụng dữ liệu huấn luyện.
4. **Điều chỉnh tham số:** Sử dụng cross-validation để chọn tham số tối ưu.
5. **Đánh giá mô hình:** Sử dụng tập dữ liệu kiểm tra.

kNN

k-Nearest Neighbors (kNN) là một thuật toán dựa trên khoảng cách giữa các điểm dữ liệu.

- **Các bước hiện thực:**
  1. **Chuẩn bị dữ liệu:** Chuẩn hóa dữ liệu.
  2. **Chọn số k:** Xác định số lượng láng giềng gần nhất.
  3. **Xây dựng mô hình:** Lưu trữ dữ liệu huấn luyện.
  4. **Dự đoán:** Tính khoảng cách đến các láng giềng và dự đoán.

Random Forest

Random Forest là một tập hợp các cây quyết định, giúp cải thiện độ chính xác và giảm overfitting.

- **Các bước hiện thực:**
  1. **Chuẩn bị dữ liệu:** Xử lý dữ liệu bị thiếu, chuẩn hóa dữ liệu.
  2. **Xây dựng mô hình:** Tạo nhiều cây quyết định từ các tập con của dữ liệu.
  3. **Huấn luyện mô hình:** Huấn luyện từng cây quyết định.
  4. **Đánh giá mô hình:** Tính trung bình dự đoán từ tất cả các cây.

Các thuật toán Regression

Các thuật toán hồi quy như Linear Regression, Ridge Regression, Lasso Regression được sử dụng để dự đoán giá trị liên tục.

- **Các bước hiện thực:**
  1. **Chuẩn bị dữ liệu:** Xử lý dữ liệu bị thiếu, chuẩn hóa dữ liệu.
  2. **Xây dựng mô hình:** Chọn mô hình hồi quy thích hợp.
  3. **Huấn luyện mô hình:** Sử dụng dữ liệu huấn luyện để tìm tham số.
  4. **Đánh giá mô hình:** Sử dụng dữ liệu kiểm tra để đánh giá.

4. Thực nghiệm

4.1. Dữ liệu

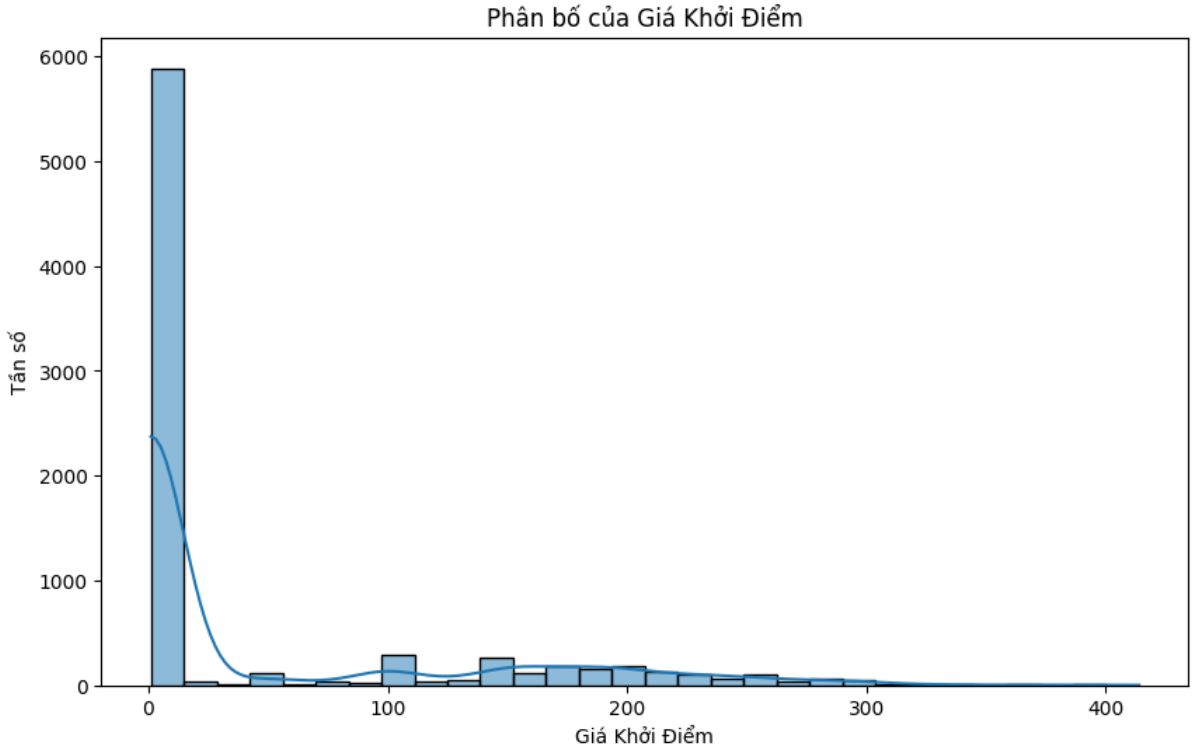
4.1.1. Mô tả dữ liệu

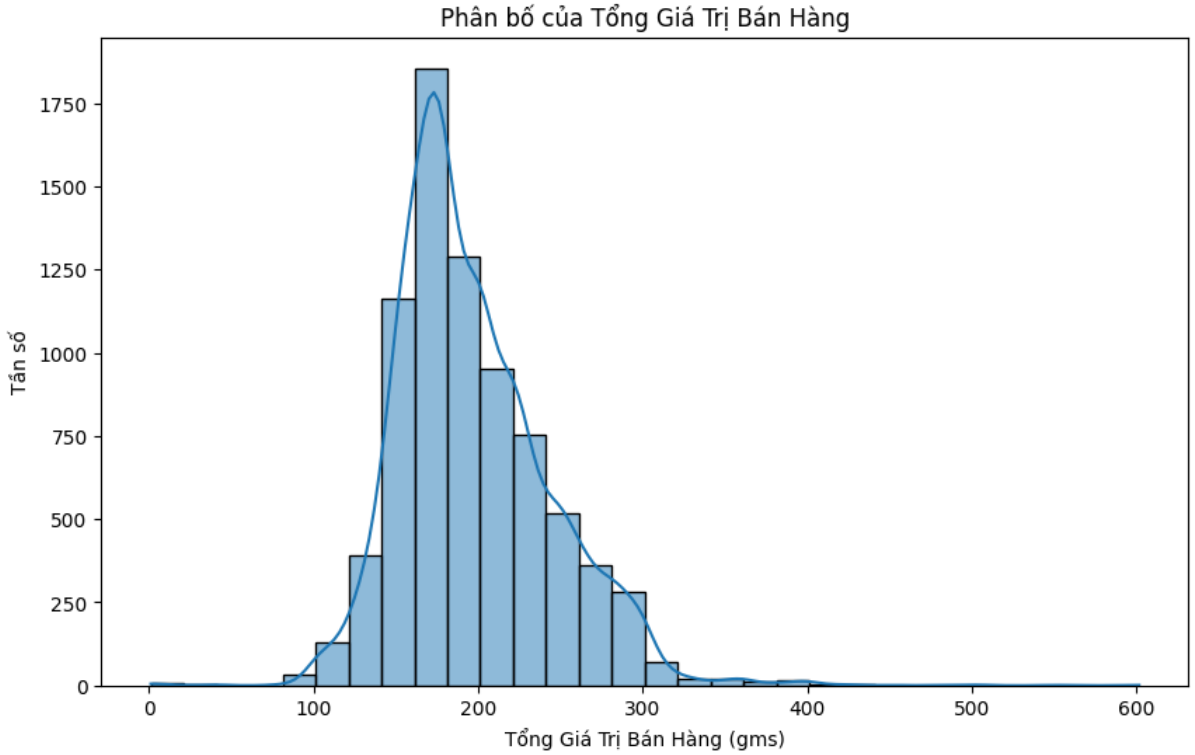
Dữ liệu bao gồm các đặc điểm của phiên đấu giá và doanh thu bán hàng. Các thuộc tính chính bao gồm:

- auct_id(int): ID của phiên đấu giá
- item_leaf_category_name (string): Tên danh mục sản phẩm
- listing_title (string): Tiêu đề sản phẩm
- listing_subtitle (string): Phụ đề sản phẩm
- listing_start_date (datetime): Ngày bắt đầu phiên đấu giá
- listing_end_date (datetime): Ngày kết thúc phiên đấu giá
- listing_durtn_days (int): Thời gian đấu giá
- listing_type_code (int): Mã loại hình đấu giá
- feedback_score_at_listing_time (int): Điểm phản hồi tại thời điểm đấu giá
- start_price (float): Giá khởi điểm
- buy_it_now_price (float): Giá mua ngay
- buy_it_now_listed_flag (bool): Cờ giá mua ngay
- bold_fee_flag, featured_fee_flag, category_featured_fee_flag, gallery_fee_flag, gallery_featured_fee_flag, ipix_featured_fee_flag, reserve_fee_flag, highlight_fee_flag, schedule_fee_flag, border_fee_flag (bool): Các cờ phí dịch vụ
- qty_available_per_listing (int): Số lượng sản phẩm có sẵn trong mỗi phiên đấu giá
- gms (float): Doanh thu bán hàng
- category_avg_gms (float): Doanh thu trung bình của danh mục sản phẩm
- gms_greater_avg (bool): Cờ doanh thu cao hơn trung bình

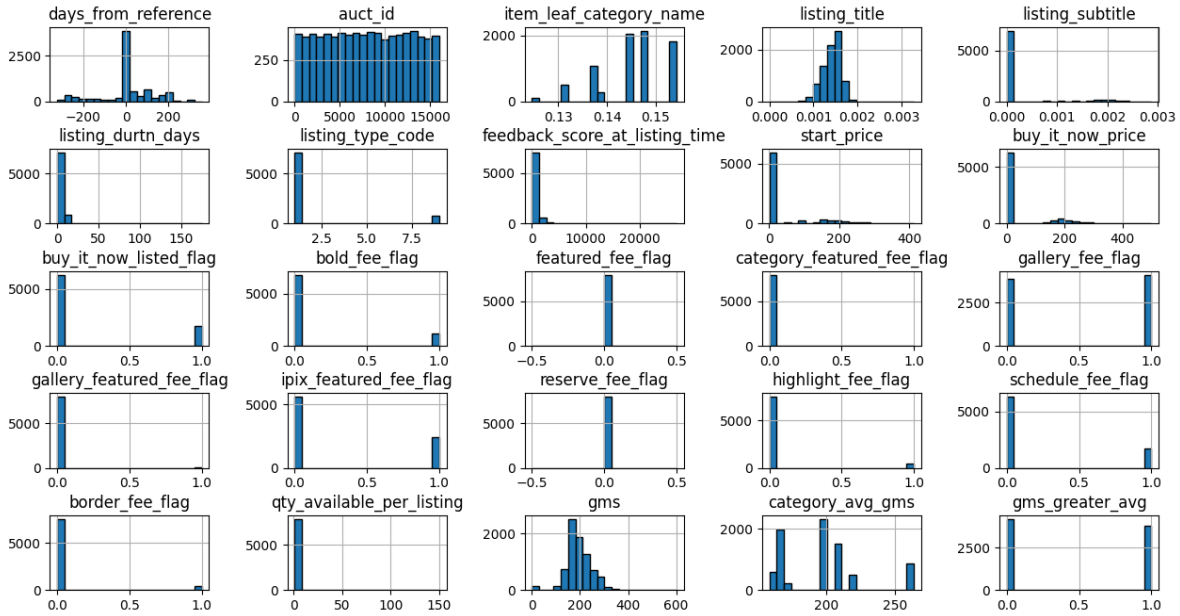
4.1.2. Biểu đồ phân bố của dữ liệu

Hình 1 . Biểu đồ tần số của Giá Khởi Điểm





Hình 2 .Biểu đồ tần số của Tổng giá trị bán hàng

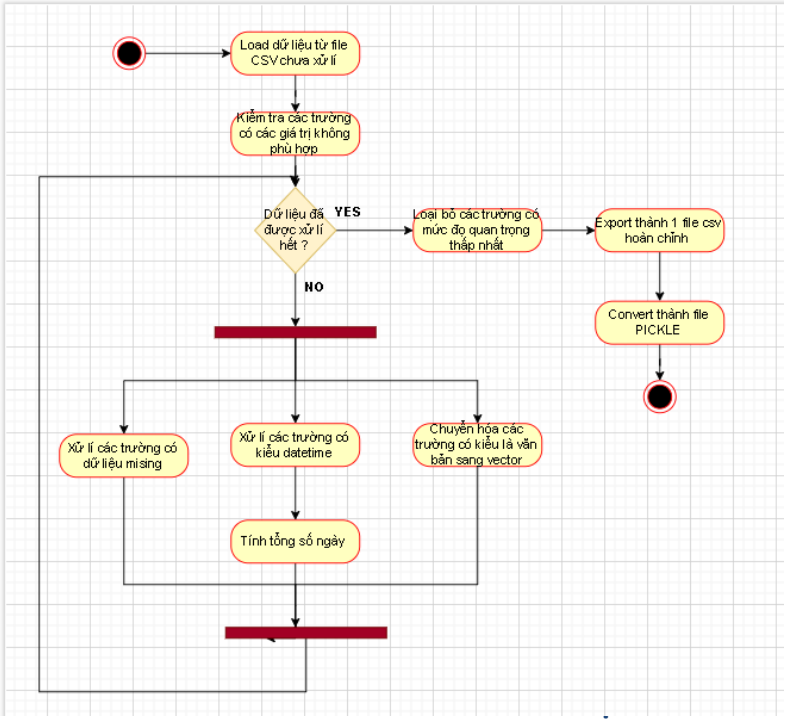


Hình 3 .Tổng quan tần số của các Thuộc tính

4.1.3. Tiền xử lí dữ liệu

- **Xử lý dữ liệu missing**: điền giá trị trung bình cho các giá trị missing giúp đảm bảo rằng dữ liệu không có lỗ hổng và vẫn giữ được tính thống kê của các thuộc tính.
- **Chọn các thuộc tính có ảnh hưởng đến dự đoán doanh số như** : *item_leaf_category_name,listing_durtn_days,feedback_score_at_listing_time,start_price,buy_it_now_price,qty_available_per_listing,bold_fee_flag, featured_fee_flag, category_featured_fee_flag, gallery_fee_flag,gallery_featured_fee_flag,ipix_featured_fee_flag,reserve_fee_flag,highlight_fee_flag,schedule_fee_flag,border_fee_flag* giảm độ phức tạp của mô hình, tăng hiệu quả huấn luyện và cải thiện độ chính xác của mô hình.
- **Rời rạc hóa một số thuộc tính liên tục như** : start_price, buy_it_now_price, feedback_score_at_listing_time biến các thuộc tính liên tục thành các nhóm giá trị rời rạc, dễ xử lý hơn trong một mô hình học máy và có thể cải thiện hiệu suất của các thuật toán.
- **Xử lí các trường dữ liệu** là văn bản sang vector áp dụng kĩ thuật chuẩn hóa văn bản sang vector, các trường là kiểu dữ liệu datetime .

Để áp dụng mô hình và các thuật toán phân lớp với dữ liệu trên

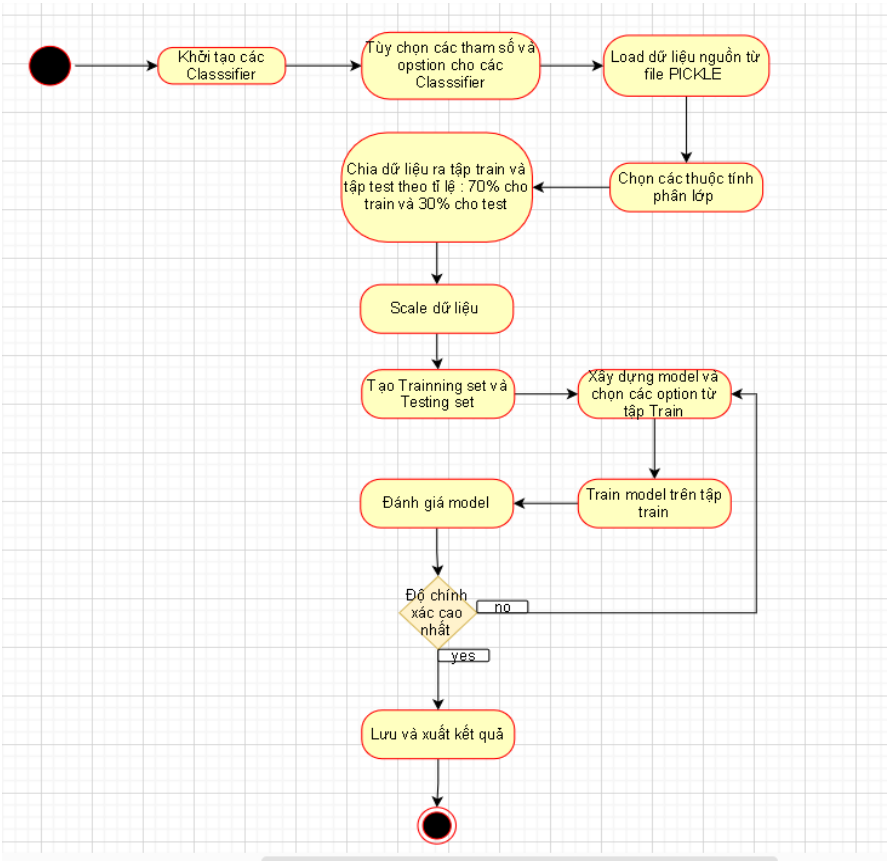


Phân chia dữ liệu

– Training set: 70%, Validation set: 15%, Testing set: 15%

4.2 Phương pháp

4.2.1 Kết quả



Hình 5 . Mô hình áp dụng

Kết quả so sánh giữa các thuật toán dựa trên tập train và tập test và mô hình áp dụng ở trên :

Bảng 1 .Kết Quả Tỉ Lệ Mô Hình

Thuật toán	Tỉ lệ dự đoán đúng	Precission	Recall	F1
SVM	99.25%	99.25%	99.25%	99.25%
KNN	94.83%	94.83%	94.83%	94.83%
Neural	63.3%	61.37%	60.06%	61.02%
Randomforest	85.3%	85.3%	85.3%	85.3%
Regression	99.8%	99.8%	99.8%	99.8%

- Dựa vào bảng trên ta có thể thấy rằng SVM cho thấy hiệu suất rất cao và ổn định trên tất cả các tiêu chí với tỉ lệ dự đoán đúng là 99.25%, tuy nhiên có thể không hiệu quả với các bộ dữ liệu lớn do thời gian tính toán lâu và yêu cầu bộ nhớ cao.

- Thuật toán KNN có tỉ lệ dự đoán đúng 94.83%, dễ hiểu và triển khai, nhưng có thể chậm và nhạy cảm với dữ liệu nhiễu và còn tùy thuộc vào việc chọn tham số k . Về Phần Neural Network có hiệu suất thấp nhất với tỉ lệ dự đoán đúng 63.3%, có thể do thiếu dữ liệu hoặc mô hình chưa được tối ưu, mặc dù có khả năng học hỏi các mẫu phức tạp.
- Random Forest đạt tỉ lệ dự đoán đúng 85.3%, tốt trong xử lý dữ liệu không cân bằng và giảm thiểu overfitting, nhưng vẫn không bằng SVM và Regression.
- Thuật toán Regression là mô hình có hiệu suất cao nhất với tỉ lệ dự đoán đúng 99.8%, dễ triển khai và tính toán nhanh chóng, nhưng có thể bị hạn chế nếu dữ liệu không tuyến tính. Tổng kết lại, Regression và SVM là hai mô hình có hiệu suất tốt nhất, trong khi KNN và Random Forest cũng cho kết quả khá tốt. Neural Network cần được tối ưu hóa thêm để cải thiện hiệu suất.

5. Kết luận

- Báo cáo đã cho ta thấy được mức độ hiệu quả của các thuật toán dựa trên các mô hình để dựa vào đó ta có thể đánh giá được mức độ và hiệu quả đạt được với các tỉ lệ mà mô hình đã đạt được từ đó cho ta thấy những mô hình và thuật toán cần được cải thiện trong tương lai cùng với dữ liệu để cho ra kết quả đạt được là tốt nhất .
  Những đánh giá ưu nhược điểm của báo cáo :
- Ưu điểm :
  Dự án này đã thành công trong việc xây dựng mô hình dự đoán doanh thu bán hàng thực tế có cao hơn doanh thu trung bình của danh mục sản phẩm hay không.
  Các mô hình học máy được áp dụng đã cho thấy hiệu quả khác nhau, với một số mô hình và thuật toán như SVM và Regression đạt hiệu suất cao nhất.
- Nhược điểm :
  Báo cáo vẫn chưa thấy được những sai sót trong việc áp dụng các mô hình và thuật toán , các tham số , param cụ thể cho từng thuật toán
  Một vài các param có thể thay đổi hoàn toàn mực độ tính chính xáctrong việc dự đoán kết quả dựa trên tập train

Tài liệu tham khảo

Dataset: Prediction of Auction Sales

Predicting The Final Price of Online Auction Items(October 2006 - **Xuefeng Li**)

Auctions versus Posted Prices in Online Markets (Liran Einav, Chiara Farronato, Jonathan Levin and Neel Sundaresany - January 2016)

Bayesian Structural Time Series (January 2020 - Abdullah M Almarashi , Khushnoor Khan)

https://phuongvu.me/phan-1-gioi-thieu-ve-dau-gia/

Bảng Phân Công Công Việc :

21130382 Nguyễn Văn Huy	Thu thập dữ liệu , giới thiệu dự án , tìm kiếm các công trình liên quan , phát biểu bài toán và thuật toán liên quan , mô tả thuật toán , mô tả dữ liệu, mô tả tiền xử lí dữ liệu, vẽ biểu đồ phân bố thuộc tính của dữ liệu
21130349 Đỗ Xuân Hậu	Vẽ biểu đồ phân bố thuộc tính của dữ liệu, tiền xử lí dữ liệu, tìm kiếm các bài toán và thuật toán liên quan, mô tả tiền xử lí dữ liệu, thực nghiệm mô hình , phương pháp , mô hình áp dụng , trình bày và đánh giá kết quả , kết luận