Trường ĐH Khoa Học Tự Nhiên Tp. Hồ Chí Minh
TRUNG TÂM TIN HỌC

Đồ án tốt nghiệp Data Science

Topic 1: ***Regression & Time Series Prediction***

Phòng LT & Mạng

https://csc.edu.vn/data-science-machine-learning/Data-Science-Capstone-Project-Hinh-thuc-2_225

2021

Nội dung

1. **Giới thiệu project**

2. **Triển khai project theo Data Science Process**

Giới thiệu project

❑ Price Prediction

Giới thiệu project

❑ Business Objective/Problem

- Bơ "**Hass**", một công ty có trụ sở tại Mexico, chuyên sản xuất nhiều loại quả bơ **được bán ở Mỹ**. Họ đã rất thành công trong những năm gần đây và muốn **mở rộng**. Vì vậy, họ muốn xây dựng mô hình hợp lý để **dự đoán giá trung bình của bơ "Hass" ở Mỹ** nhằm xem xét việc **mở rộng** các loại trang trại Bơ đang có cho việc trồng bơ ở các vùng khác.

https://www.dezyre.com/project-use-case/perform-time-series-modelling-facebook-prophet

Giới thiệu project

- **Các kiến thức/ kỹ năng cần để giải quyết vấn đề này:**
  - Hiểu vấn đề
  - Import các thư viện cần thiết và hiểu cách sử dụng
  - Đọc dữ liệu (project này cung cấp dataset)
  - Thực hiện EDA cơ bản (sử dụng *Pandas Profiling Report)*
  - Tiền xử lý dữ liệu: làm sạch, tạo tính năng mới, lựa chọn tính năng cần thiết…

Giới thiệu project

- Làm việc với dữ liệu thời gian
- Trực quan hóa dữ liệu các loại
- Lựa chọn thuật toán cho bài toán regression và bài toán time series analysis
- Xây dựng model
- Đánh giá model
- Báo cáo kết quả

Nội dung

1. Giới thiệu project

2. **Triển khai project theo Data Science Process**

Triển khai project theo Data Science Process

- Thư viện sử dụng
  - Numpy, pandas, matplotlib, seaborn,
  - pandas_profiling
  - scikit-learn (sklearn), xgboost
  - pmdarima, fbprophet
  - …

Triển khai project theo Data Science Process

❑ Triển khai dự án

- Bước 1: Business Understanding
  - Dựa vào mô tả nói trên (hoặc sau khi đặt ra các câu hỏi cụ thể cho doanh nghiệp và các đối tượng có liên quan) => xác định được vấn đề:
    - Hiện tại: Công ty kinh doanh quả bơ ở rất nhiều vùng của nước Mỹ với 2 loại bơ là bơ thường và bơ hữu cơ, được đóng gói theo nhiều quy chuẩn (Small/Large/XLarge Bags), và có 3 PLU (Product Look Up) khác nhau (4046, 4225, 4770). Nhưng họ chưa có mô hình để dự đoán giá bơ cho việc mở rộng.

    => Mục tiêu/ Vấn đề: Xây dựng mô hình **dự đoán giá trung bình của bơ "Hass" ở Mỹ =>** xem xét việc mở rộng sản xuất, kinh doanh.

Triển khai project theo Data Science Process

- Bước 2: Data Understanding/ Acquire
  - Từ mục tiêu/ vấn đề đã xác định: xem xét các dữ liệu mà công ty đang có:
    - Dữ liệu được lấy trực tiếp từ máy tính tiền của các nhà bán lẻ dựa trên doanh số bán lẻ thực tế của bơ Hass.
    - Dữ liệu đại diện cho dữ liệu lấy từ máy quét bán lẻ hàng tuần cho lượng bán lẻ (National retail volume- units) và giá bơ từ tháng 4/2015 đến tháng 3/2018.
    - Giá Trung bình (Average Price) trong bảng phản ánh giá trên một đơn vị (mỗi quả bơ), ngay cả khi nhiều đơn vị (bơ) được bán trong bao.
    - Mã tra cứu sản phẩm - Product Lookup codes (PLU's) trong bảng chỉ dành cho bơ Hass, không dành cho các sản phẩm khác.

Triển khai project theo Data Science Process

- Toàn bộ dữ liệu được đổ ra và lưu trữ trong tập tin avocado.csv với 18249 record. Với các cột:
  - Date - ngày ghi nhận
  - AveragePrice – giá trung bình của một quả bơ
  - Type - conventional / organic – loại: thông thường/ hữu cơ
  - Region – vùng được bán
  - Total Volume – tổng số bơ đã bán
  - 4046 – tổng số bơ có mã PLU 4046 đã bán
  - 4225 - tổng số bơ có mã PLU 4225 đã bán
  - 4770 - tổng số bơ có mã PLU 4770 đã bán
  - Total Bags – tổng số túi đã bán
  - Small/Large/XLarge Bags – tổng số túi đã bán theo size
- Có hai loại bơ trong tập dữ liệu và một số vùng khác nhau. Điều này cho phép chúng ta thực hiện tất cả các loại phân tích cho các vùng khác nhau, hoặc phân tích toàn bộ nước Mỹ theo một trong hai loại bơ.

Triển khai project theo Data Science Process

=> Có thể tập trung giải quyết hai bài toán

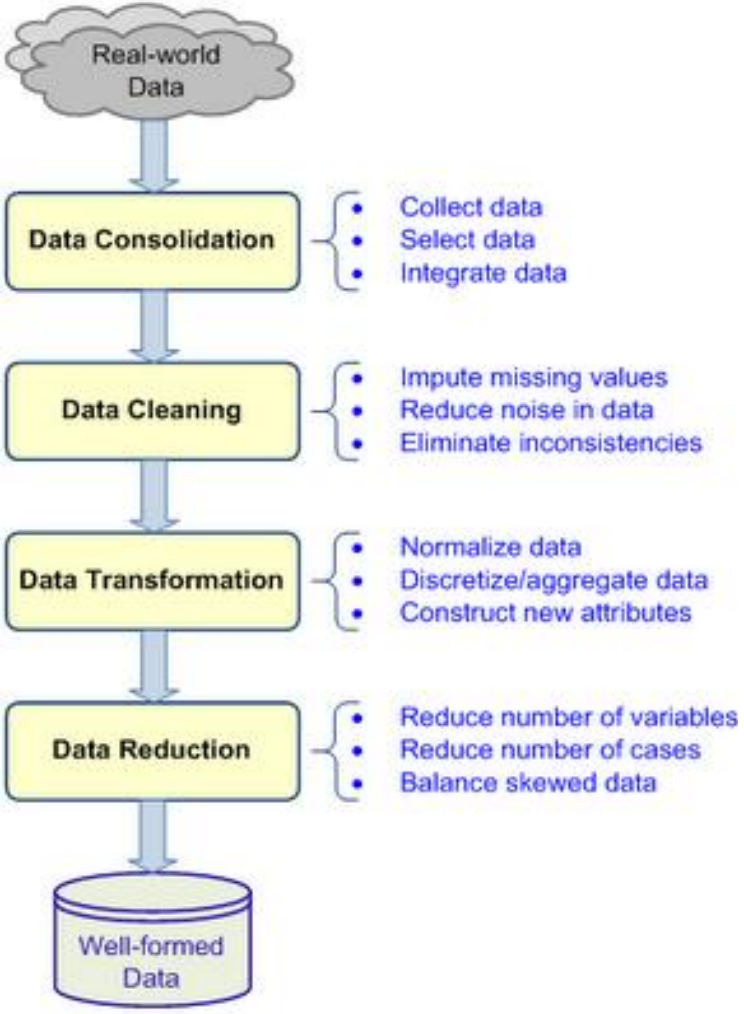
- Bài toán 1: USA's Avocado AveragePrice Prediction – Sử dụng các thuật toán Regression như Linear Regression, Random Forest, XGB Regressor...
- Bài toán 2: Conventional/Organic Avocado Average Price Prediction for the future in California/NewYork… - sử dụng các thuật toán Time Series như ARIMA, Prophet...

Triển khai project theo Data Science Process

- Bước 3: Data preparation/ Prepare
  - Từ bước 3 trở đi cách triển khai cho hai bài toán sẽ khác nhau.

Triển khai project theo Data Science Process

- Với bài toán 1: Thực hiện các công việc



Triển khai project theo Data Science Process

- Với bài toán 2: Chuẩn hóa dữ liệu theo TimeSeries
  - Chọn một vùng để thực hiện việc dự đoán (Ví dụ: California)
  - Với vùng được chọn chia thành 2 bài toán nhỏ hơn là loại bơ organic và loại bơ conventional
  - Chuẩn dữ liệu có cột Date (theo tháng năm) và AveragePrice là trung bình của các tuần trong tháng năm tương ứng.

Triển khai project theo Data Science Process

- Bước 4&5: Modeling & Evaluation/ Analyze & Report
  - Với bài toán 1:
    - Xây dựng các Regression model dự đoán giá
      - Linear Regression
      - Random Forest Regression
      - XGB Regression
      - …
    - Thực hiện/ đánh giá kết quả các Regression model
      - R-squared
      - MSE/ RMSE/ MAE
      - Kết luận

Triển khai project theo Data Science Process

- Với bài toán 2:
  - Xây dựng các TimeSeries model dự đoán giá trong tương lai
    - ARIMA
    - Facebook Prophet
    - ...
  - Thực hiện/ đánh giá kết quả các regression model
    - MSE/ RMSE/ MAE
    - Vẽ đồ thị gồm có các giá trị thực tế, giá trị dự báo, khoảng tin cậy…
    - Kết luận

Triển khai project theo Data Science Process

- Facebook Prophet
  - Dự báo giúp chúng ta trả lời câu hỏi “What will happen?”. Các nhà phân tích mong muốn đưa ra các dự báo với mức độ chính xác cần thiết để đưa ra các quyết định kinh doanh.
  - Có nhiều thuật toán, thư viện được xây dựng để hỗ trợ dự báo, trong đó có thư viện mã nguồn mở Prophet. Core Data Science team của Facebook đã tạo ra Prophet cho Python và R vào năm 2017.

Triển khai project theo Data Science Process

- Chức năng
  - Prophet được thiết kế để đưa ra dự báo cho các bộ dữ liệu chuỗi thời gian đơn biến.
  - Nó rất dễ sử dụng và được thiết kế để tự động tìm một bộ siêu tham số phù hợp cho mô hình trong việc đưa ra các dự báo cho dữ liệu có xu hướng (trend) và cấu trúc theo mùa (seasonal) theo mặc định.

Triển khai project theo Data Science Process

- Xây dựng model
  - Dữ liệu đầu vào của Prophet là một dataframe (ví dụ đặt tên là df) có hai cột ds và y. Cột ds (datestamp) phải theo định dạng date YYYY-MM-DD hoặc timestamp YYYY-MM-DD HH:MM:SS của pandas. Cột y phải là numeric, và đại diện cho phép đo mà chúng ta muốn dự đoán.
  - Khai báo và build model với model = Prophet() và model.fit(df)
  - Tạo future datafarme mới để dự đoán: future = model.make_future_dataframe(periods=12)
  - Dự đoán với: forecast = model.predict(future)
  - Xem kết quả: forecast[['ds', 'yhat', 'yhat_lower', 'yhat_upper']].tail()
  - Trực quan hóa kết quả: model.plot(forecast)
  - Trực quan hóa các component (trend, weekly, yearly): model.plot_components(forecast)

Demo

Triển khai project theo Data Science Process

- Bước 6: Deployment & Feedback/ Act
  - Đưa ra chiến lược kinh doanh phù hợp cho các vùng khác nhau để tăng doanh thu, mở rộng sản xuất.

Triển khai project theo Data Science Process

Các công việc cần thực hiện…

Triển khai project theo Data Science Process

Với project phía trên

- Yêu cầu 1: Bài toán 1 (0.5 điểm)
  - Thực hiện các tiền xử lý dữ liệu bổ sung (nếu cần)
  - Ngoài những thuật toán regression đã được thực hiện, có thuật toán nào khác cho kết quả tốt hơn không? Thực hiện với thuật toán đó. Tổng hợp kết quả thu được.

Triển khai project theo Data Science Process

- Yêu cầu 2: Bài toán 2 (0.5 điểm)
  - Thực hiện các thuật toán ARIMA, Facebook Prophet... để dự đoán giá, khả năng mở rộng trong tương lai.
  - Tổng hợp kết quả thu được.

Triển khai project theo Data Science Process

- Yêu cầu 3: (0.5 điểm)
  - Hãy làm tiếp phần dự đoán giá bơ thường (Conventional Avocado) của vùng California. Tổng hợp kết quả thu được.

Triển khai project theo Data Science Process

- Yêu cầu 4: (0.5 điểm)
  - Hãy chọn ra một vùng (trong danh sách các vùng hãng bơ “Hass” đang kinh doanh) mà bạn cho rằng trong tương lai có thể mở rộng trồng trọt, sản xuất và kinh doanh (Organic và/ hoặc Conventional Avocado). Hãy chứng minh điều này bằng cách triển khai các bài toán như đã làm với vùng California.

Triển khai project theo Data Science Process

Giới thiệu Lazy Predict

- Lazy Predict giúp xây dựng rất nhiều mô hình cơ bản mà không cần viết nhiều code, giúp chúng ta dễ dàng mô hình nào hoạt động tốt mà không cần bất kỳ điều chỉnh tham số nào.
- Cài đặt: pip install lazypredict
- https://pypi.org/project/lazypredict/

demo