**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

# PHÂN TÍCH NHỮNG YẾU TỐ ẢNH HƯỞNG TỚI GIÁ TIỀN SAU MỖI CHUYẾN ĐI BẰNG YELLOW TAXI


| Sinh viên thực hiện: | | |
|---|---|---|
| STT | Họ tên | MSSV |
| 1 | Nguyễn Thị Bảo Hân | 19520071 |
| 2 | Văn Kim Ngân | 19520177 |
| 3 | Tiêu Kim Hảo | 19521480 |

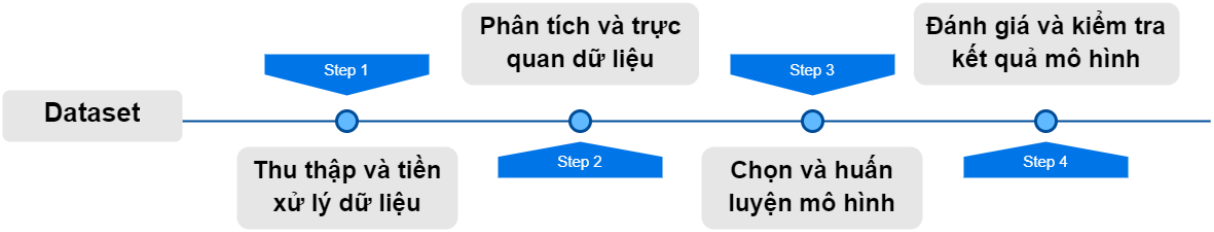
# TP. HỒ CHÍ MINH – 12/2021

## 1. GIỚI THIỆU

Đề tài này nhằm dự đoán giá tiền sau mỗi chuyến đi bằng Yellow Taxi dựa trên những chuyến đi đã được ghi lại, điều này giúp cho hãng xe có thể nắm bắt để điều phối lượng xe thích hợp để đáp ứng đủ nhu cầu của khách hàng cũng như giúp cánh tài xế có thể dự đoán số tiền mình có thể nhận sau mỗi chuyến xe. Để làm được điều này, dựa trên dữ liệu từ những chuyến đi, chúng ta sẽ phân tích tất các yếu tố để chọn ra những thuộc tính nào có ảnh hưởng đến tiền xe nhất. Sau đó những thuộc tính này sẽ được đưa vào mô hình huấn luyện để dự đoán doanh thu của hãng xe.

Bước đầu, ta cần phải nắm bắt cũng như hiểu rõ data, điều này rất quan trọng vì chỉ khi hiểu rõ data thì mới có thể phân tích một cách tốt nhất, hiệu quả nhất. Vì vậy, ta sẽ tiến hành phân tích thăm dò dữ liệu (EDA), từ đó chọn ra các yếu tố tương quan và ảnh hưởng đến tiền xe nhất. Sau khi có các thuộc tính, ta sẽ chọn nhiều mô hình để huấn luyện, so sánh kết quả giữa các mô hình nhằm chọn ra mô hình có kết quả tốt nhất. Ở đây ta sẽ sử dụng các thuật toán Random Forest và Neural Network. Qua quá trình thực nghiệm, các mô hình hoạt động khá ổn trên bộ dữ liệu, cho kết quả của độ đo RMSE khá ổn.

## 2. NỘI DUNG




*Hình 1. Quy trình Phân tích Dữ liệu.*

### 2.1. Mô tả bộ dữ liệu

Tên bộ dữ liệu: 2020 Yellow Taxi Trip Data.

Nguồn gốc: Trích từ bộ dữ liệu 2020 Yellow Taxi Trip Data trên trang NYC Open Data.

Bộ dữ liệu có 18 thuộc tính.

| Column Name | Column Description | Term, Acronym, or Code Definitions |
|---|---|---|
| VendorID | Mã cho biết nhà cung cấp dịch vụ công nghệ taxicab (TPEP) đã cung cấp hồ sơ | 1= Creative Mobile Technologies, LLC;<br>2= Curb Mobility (trước đây là VeriFone Inc) |
| tpep_pickup_datetime | Thời gian ngày và giờ được tính từ khi đồng hồ đo trong taxi hoạt động. | |
| tpep_dropoff_datetime | Thời gian ngày và giờ được tính từ khi đồng hồ đo trong taxi dừng. | |

| | | |
|---|---|---|
| Passenger_count | Số lượng hành khách trên xe | |
| Trip_distance | Khoảng cách di chuyển được tính theo đồng hồ đo. | |
| PULocationID | Địa điểm đón khách, giá trị dữ liệu là mã số vùng được lấy từ TLC Taxi Zone | Mã từ 0-263 tương ứng với từng khu vực |
| DOLocationID | Địa điểm trả khách, giá trị dữ liệu là mã số vùng được lấy từ TLC Taxi Zone | Mã từ 0-263 tương ứng với từng khu vực |
| RateCodeID | Mã cước cho từng loại chuyến đi taxi | 1=Di chuyển bình thường<br>2=Di chuyển tới sân bay JFK<br>3= Di chuyển tới sân bay Newark<br>4=Di chuyển tới quận Nassau hoặc Westchester<br>5=Giá cả thỏa thuận<br>6=Group ride |
| Store_and_fwd_flag | Thuộc tính này cho biết liệu bản ghi chuyến đi có được lưu trong bộ nhớ xe trước khi gửi đến nhà cung cấp hay không, còn gọi là "lưu trữ và chuyển tiếp", vì xe không có kết nối với máy chủ. | Y= chuyến đi đã được lưu trữu và chuyển tiếp<br>N= chuyến đi không được lưu trữu và chuyển tiếp |
| Payment_type | Phương thức thanh toán | 1= Trả thẻ<br>2= Tiền mặt<br>3=Không thanh toán<br>4=Tranh chấp<br>5=Không xác định<br>6=Chuyến đi bị hoãn |
| Fare_amount | Giá tiền tính theo công tơ mét | |
| Extra | Tiền phụ phí, hiện tại, mức phí này chỉ bao gồm 0,5 đô và 1 đô cho giờ cao điểm và phí qua đêm | |
| MTA_tax | Thuế di chuyển bằng phương tiện đi lại trong đô thị, với mức giá là 0.5 đô | |
| Improvement_surcharge | Tiền phụ phí cải thiện trên mỗi chuyến đi, giá là 0.3 đô | |
| congestion_surcharge | Phụ phí tắc nghẽn, được bổ sung vào giá cước khi chuyến đi bị tắc nghẽn | |
| Tip_amount | Tiền khách tặng thêm cho tài xế, không bao gồm tiền mặt | |
| Tolls_amount | Tổng số tiền phí cầu đường mà khách hàng phải thanh toán. | |
| Total_amount | Tổng số tiền khách hàng trả sau mỗi chuyến đi. Không bao gồm tip bằng tiền mặt. | |

*Bảng 1. Mô tả bộ dữ liệu.*

## 2.2. Thu thập và tiền xử lý dữ liệu

Dữ liệu được nhóm thu thập trực tiếp từ trang NYC Open Data. Bộ dataset này không bị khuyết dữ liệu. Dựa trên mô tả của bộ dữ liệu, ta có thể xác định được các thuộc tính của biến liên tục và biến phân loại. Tất cả các thuộc tính của dữ liệu khi vừa được thu thập về đều ở dạng 'object' nên chúng ta cần phải xem xét và điều chỉnh cho các thuộc tính về đúng kiểu dữ liệu của nó để có thể phân tích. Ở dây ta có thể thấy các thuộc tính liên tục như *['trip_distance', 'fare_amount', 'tip_amount', 'tolls_amount', 'total_amount']* đang ở sai kiểu dữ liệu cần được astype về dạng 'float'. 2 thuộc tính *['tpep_pickup_datetime', 'tpep_dropoff_datetime']* là những thuộc tính chỉ thời gian nên nhóm đưa về kiểu dữ liệu 'datetime64[ns]'.

Thuộc tính *['Duration']* mang ý nghĩa là khoảng thời gian hành khách ngồi trên xe. Thuộc tính này được nhóm bổ sung vào bằng cách lấy thời gian hành khách xuống xe *['tpep_dropoff_datetime']* trừ đi cho thời gian bắt đầu lên xe *['tpep_pickup_datetime']*, thời gian này được tính bằng phút.

Thuộc tính *['Time']* được tách ra từ thuộc tính *['tpep_pickup_datetime']* là thời điểm giờ mà hành khách lên xe, dựa trên định dạng 24h.

Ta có thể thấy dữ liệu của thuộc tính *['tpep_pickup_datetime']* đã được khai thác thành những thuộc tính riêng biệt, việc giữ nguyên dữ liệu của thuộc tính hiện giờ không còn ý nghĩa trong việc phân tích nên nhóm đã quyết định chuyển đổi dữ liệu từ định dạng ngày/tháng/năm về dạng thứ trong tuần. Với thứ hai, thứ ba, thứ tư, thứ năm, thứ sáu, thứ bảy, chủ nhật lần lượt là mon, tues, wed, thu, fri, sat, sun. Ngoài ra ta có thể loại bỏ thuộc tính *['tpep_dropoff_datetime']* ra khỏi dataset vì thuộc tính đã được khai thác.

Trong quá trình xử lý dữ liệu, nhóm phát hiện rằng thuộc tính *['passenger_count']* có những điểm dữ liệu mang giá trị lớn hơn số lượng người tối đa mà taxi có thể chứa. Ngoài ra nhóm cũng loại bỏ sample mang giá trị âm trong 2 thuộc tính như *['trip_distance'], ['total_amount'].* Những điểm dữ liệu này có số lượng ít nên chúng ta sẽ drop những điểm dữ liệu này.

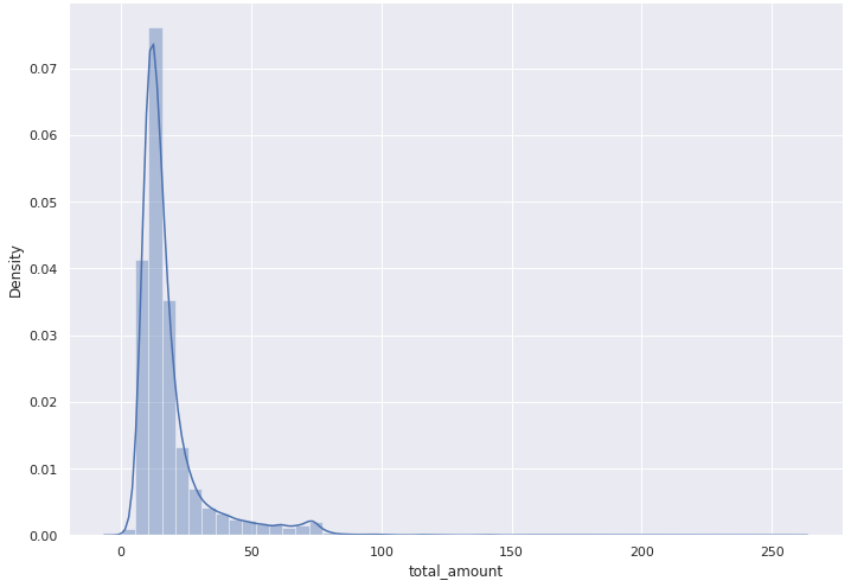
## 2.3. Phân tích dữ liệu

**Thống kê mô tả dữ liệu:**

| | trip_distance | fare_amount | tip_amount | tolls_amount | total_amount | duration |
|---|---|---|---|---|---|---|
| **count** | 986174 | 986174 | 986174 | 986174 | 986174 | 986174 |
| **mean** | 3.151618 | 12.79267 | 2.067237 | 0.37663 | 18.62534 | 15.82132 |
| **std** | 4.153618 | 12.19606 | 2.785432 | 1.921701 | 14.93499 | 68.63807 |
| **min** | 0.01 | 0 | 0 | 0 | 0.3 | 0.016667 |
| **25%** | 1 | 6 | 0 | 0 | 10.8 | 6.033333 |
| **50%** | 1.69 | 9 | 1.76 | 0 | 13.8 | 9.916667 |
| **75%** | 3.18 | 13.5 | 2.75 | 0 | 19.55 | 16 |
| **max** | 259.22 | 1238 | 450 | 910.5 | 1242.3 | 2458.4 |

*Bảng 2. Thống kê mô tả bộ dữ liệu.*

**Hình dạng của dữ liệu:**

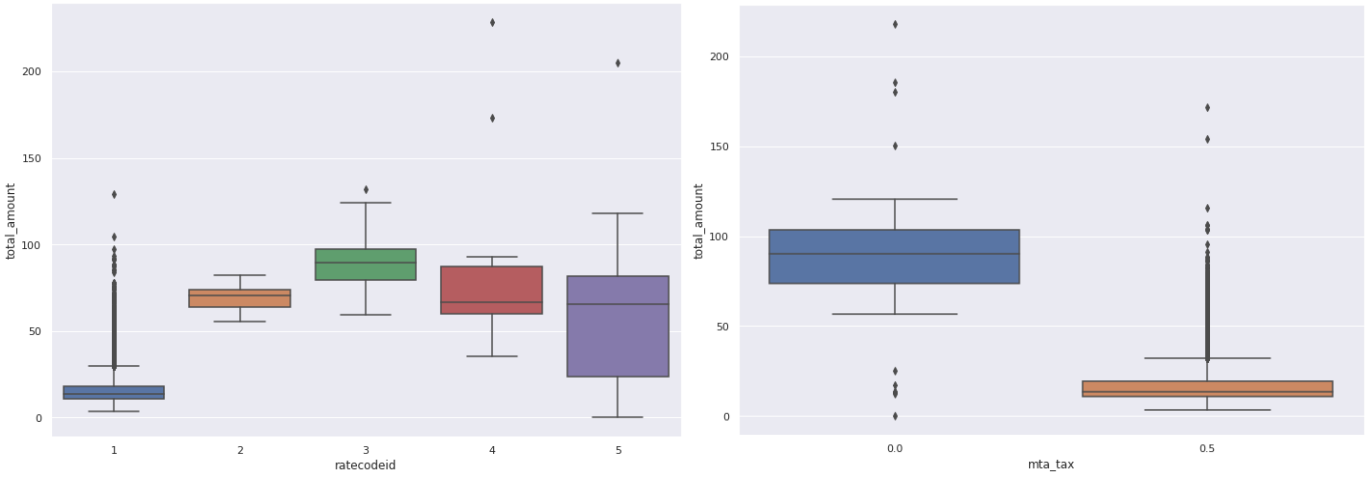



*Hình 2. Distribution của total_amount.*

Nhận xét:

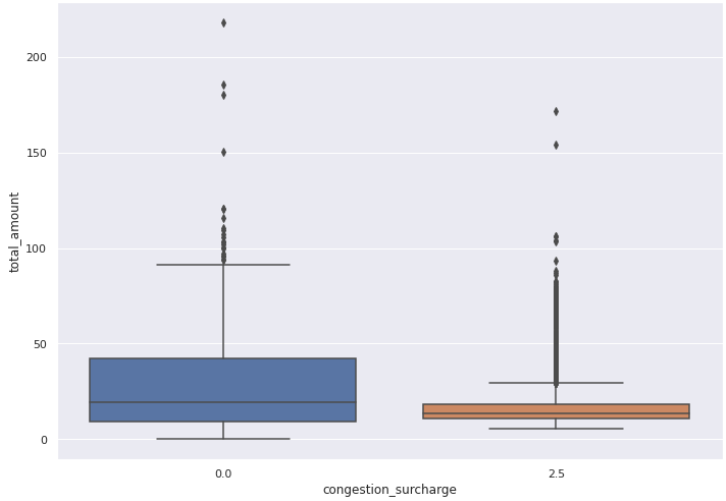
- Distribution có giá trị skew là 4.6377764096732506 → ta có thể thấy rằng dữ liệu đang bị lệch phải.
- Distribution có giá trị kurtosis là 97.54306324166436 → kurtosis càng cao thì càng có nhiều ngoại lệ, vì vậy dữ liệu này khá nhiều giá trị ngoại lệ.

**Boxplot giữa total_amount với các thuộc tính có ảnh hưởng:**




Nhận xét:

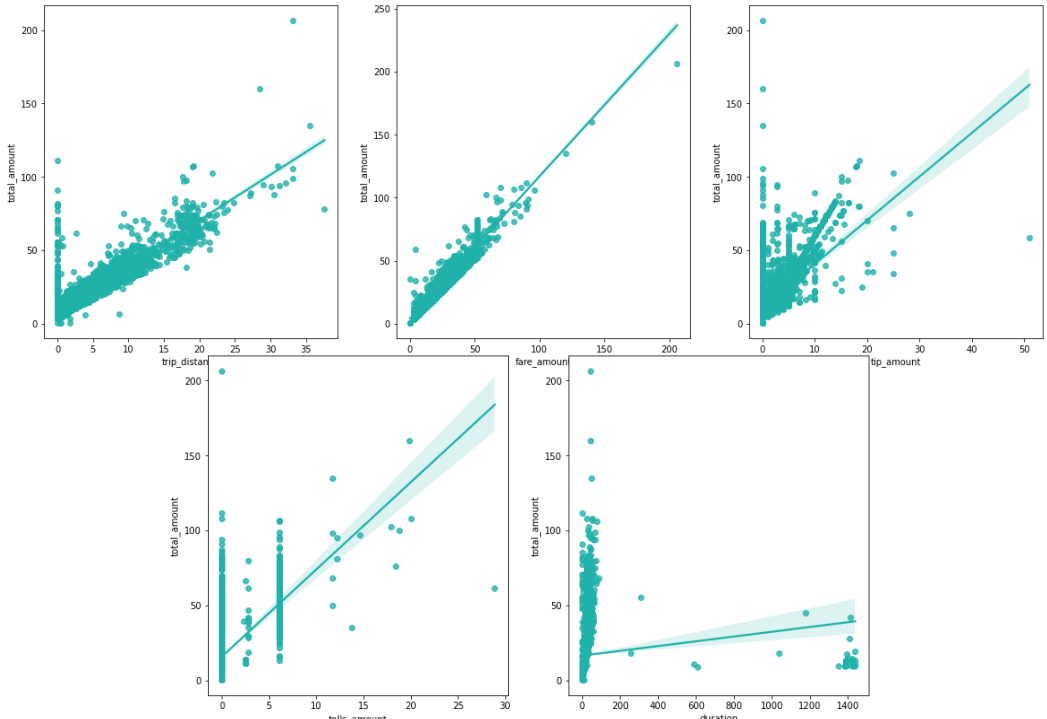
- Nhìn vào ta có thể thấy ratecodeid có ảnh hưởng một phần tới total_amount vì các giá trị có sự chênh lệch với nhau.
- Mta_tax thấy rõ sự phân bố chênh lệch hoàn toàn giữa hai giá trị 0.0 và 0.5. Vì vậy mta_tax có ảnh hưởng đến total_amount.




*Hình 3. Boxplot total_amount với yếu tố.*

- Đối với congestion_surcharge, giá trị 2.5 tập trung ở một khoảng giá trị rất hẹp và thấp, còn giá trị 0.0 phân bố rộng và có total_amount cao.

**Biểu đồ hồi quy tuyến tính:**

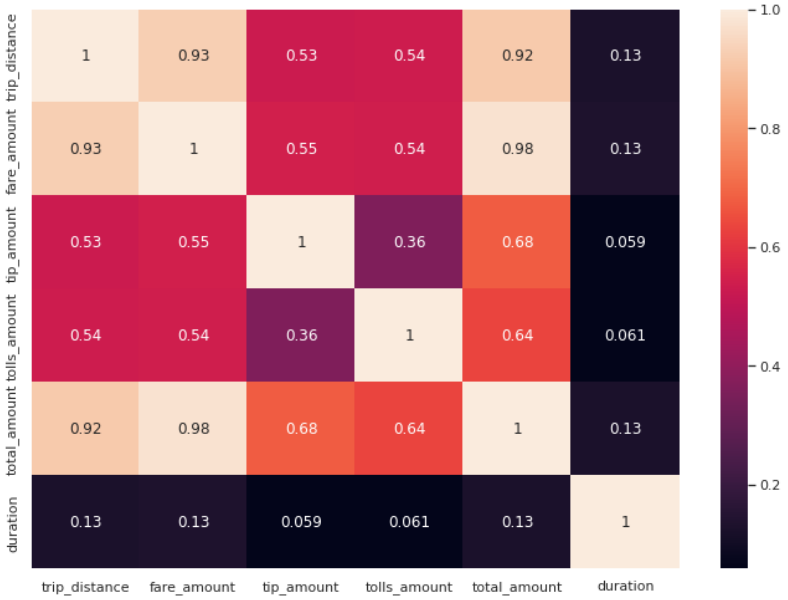



*Hình 4. Regression plot của các thuộc tính.*

Nhận xét:

- Dựa vào Regression plot này, ta có thể thấy trip_distance, fare_amount và tip_amount phân bố rải rác trên đường hồi quy. Nên có thể kết luận ba thuộc tính trên có tương quan thuận đối với total_amount.

**Ma trận tương quan:**




*Hình 5. Correlation Matrix.*

Nhận xét:

- Dựa vào ma trận tương quan, ta có thể xác định các thuộc tính có tương quan với total_amount là: trip_distance, fare_amount, tip_amount, tolls_amount.

**Chạy thử nghiệm ANOVA kiểm tra tính ảnh hưởng của các biến liên tục:**

Nếu chúng ta bác bỏ giả thuyết $H_0$, mô hình có ý nghĩa đối với ít nhất một biến giải thích. Nếu chúng ta không bác bỏ $H_0$, mô hình không có ý nghĩa đối với ít nhất một biến. Thống kê F-test sẽ cho chúng ta biết liệu mô hình của chúng ta trên thực tế có ý nghĩa hay không và sẽ tuân theo phân phối F.

```
Analysis of Variance Table

Response: total_amount
               Df  Sum Sq Mean Sq    F value Pr(>F)
trip_distance   1 1766979 1766979 2.4027e+06 <2e-16 ***
fare_amount     1  430291  430291 5.8511e+05 <2e-16 ***
tip_amount      1   63553   63553 8.6419e+04 <2e-16 ***
tolls_amount    1   16600   16600 2.2572e+04 <2e-16 ***
duration        1       1       1 8.6360e-01 0.3528
Residuals    9994    7350       1
```

*Hình 6. Bảng báo cáo thử nghiệm ANOVA.*

P-value của F-test cho tương tác giữa các thuộc tính sau:

- Duration: 0.3528 > 0.05 → Vì vậy không thể bác bỏ $H_0$ (tức là kết luận không có tương tác giữa các factor ở mức ý nghĩa 0,05) → Loại bỏ duration.

**Chạy thử nghiệm t – test:**

```
lm(formula = total_amount ~ trip_distance + fare_amount + tip_amount +
    tolls_amount, data = tam)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9553 -0.1870 -0.0594  0.3787  6.6602

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.520456   0.013325 264.193  < 2e-16 ***
trip_distance -0.023061   0.004551  -5.067 4.12e-07 ***
fare_amount    0.988286   0.001409 701.269  < 2e-16 ***
tip_amount     1.044643   0.003791 275.567  < 2e-16 ***
tolls_amount   1.016376   0.006765 150.241  < 2e-16 ***
```

*Hình 7. Summary Statistics Tables.*

Nhận xét:

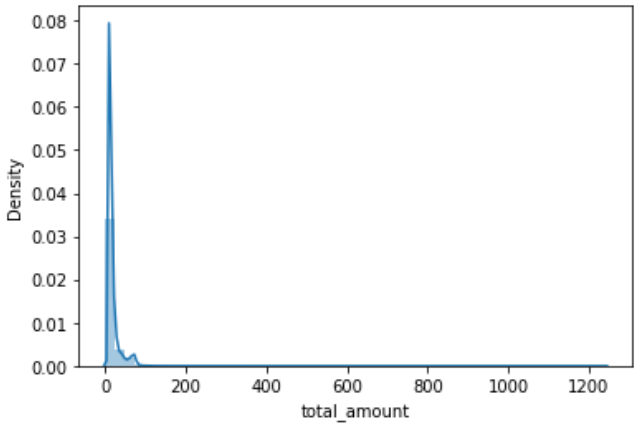
- Tất cả các thuộc tính trên đều có p-value < 0.05 → Vì vậy bác bỏ $H_0$.
- Sau khi chạy thử nghiệm t-test thì các thuộc tính trên là những thuộc tính quan trọng trong các biến liên tục, có ảnh hưởng tới mô hình.

**Kết luận:** Sau khi tiến hành phân tích dữ liệu bằng cách visualize dữ liệu và kiểm tra cả ANOVA, t – test. Ta thu được các biến được cho là quan trọng: trip_distance, fare_amount, tip_amount, tolls_amount, mta_max, congestion_surcharge, ratecode_id.

## 2.4. Huấn luyện mô hình

Chúng tôi sử dụng hai thuật toán là Random Forest và Neural Network để huấn luyện cho mô hình.

Trong quá trình huấn luyện mô hình, chúng tôi thấy dữ liệu trên biến phụ thuộc phân bố chủ yếu trong khoảng 0-100 và số có khoảng 2000 điểm dữ liệu từ 100-12000, con số này quá nhỏ so với 1 triệu điểm dữ liệu của mô hình. Tuy nhiên nó ảnh hưởng rất nhiều đến kết quả của các mô hình, nên chúng tôi chỉ giữ lại các giá trị từ 0-100 để huấn luyện mô hình.




*Hình 8. Distribution của total_amount.*

**Random Forest:**

Random Forest là mô hình tính giá trị trung bình các giá trị dữ đoán của cây quyết định, các cây quyết định được tạo ra bằng cách thay đổi ngẫu nhiên các tham số khác nhau để chỉ định dữ liệu, tham số nào được sử dụng cho huấn luyện.

Trong Random Forest có khoảng 5 cách để xử lý các biến phân loại như trong bài bài báo "Splitting on categorical predictors in random forest" đã trình bày .

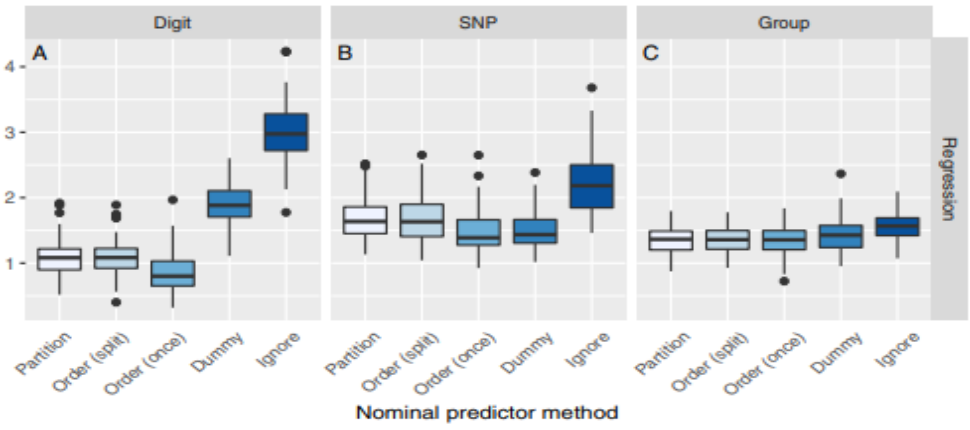



Table 3 **Median runtimes of simulations (in seconds).**

| Predictor type | Outcome type | Nominal predictor method | | | | |
|---|---|---|---|---|---|---|
| | | Partition | Order (split) | Order (once) | Dummy | Ignore |
| Digit | Regression | 15.89 | 6.22 | **4.69** | 6.60 | 4.93 |
| | Binary class. | 18.23 | 4.27 | **3.80** | 5.92 | 5.62 |
| | Multiclass class. | 28.47 | 11.09 | **5.77** | 13.19 | 9.84 |
| | Survival | 660.43 | 74.85 | 61.88 | **9.24** | 51.45 |

*Hình 9. Median runtimes of simulations.*

Ta thấy phương pháp Order(once) (sắp xếp thứ tự trước khi huấn luyện) cho biến phân loại thuộc kiểu Digit trong bài toán regression cho độ lỗi và thời gian thấp nhất. Vì thế chúng tôi sử dụng phương pháp này để huấn luyện mô hình. Do các giá trị trong biến phân loại không phân biệt thứ tự nên chúng tôi không sắp xếp trước khi huấn luyện mô hình.

Chúng tôi sử dụng class TabularPandas trong thư viện Fastai để chuyển giá trị của các biến phân loại về dạng category và chia ngẫu nhiên 80% cho tập train và 20% cho tập valid.

Một trong những đặc tính quan trọng nhất của random forest là nó không quá nhạy cảm với các lựa chọn siêu tham số. mô hình càng nhiều cây quyết định, độ chính xác càng cao.

Chúng tôi sử dụng bộ tham số n_estimators = 30, max_samples = 200_000, max_features = 0.5, min_samples_leaf = 4 và độ đo là RMSE để huấn luyện mô hình. Trong đó, n_estimators là số lượng cây quyết định, max_samples và max_features là số lượng hàng và số lượng cột sử dụng để huấn luyện mỗi cây, min_samples_leaf là giá trị khi số lượng samples trong nodes lá bằng với giá trị này thì sẽ không chia nhánh nữa.

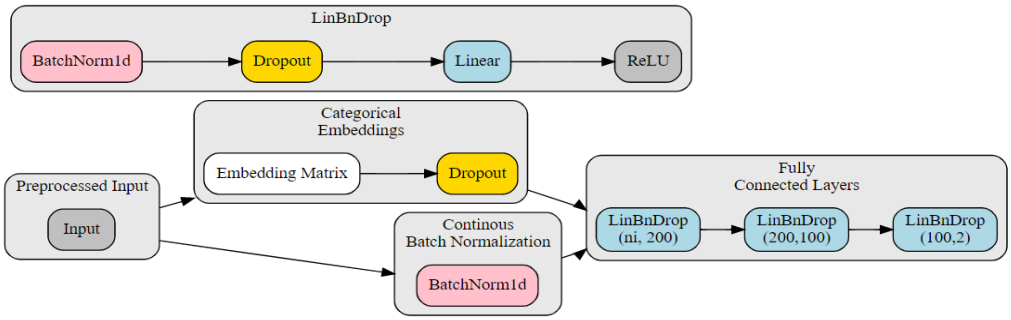
Chúng tôi thực hiện huấn luyện mô hình trên bộ dữ liệu trước và sau khi đã loại bỏ các yếu tố ảnh hưởng thấp tới mô hình để cho thấy độ hiệu quả của việc EDA.

| | | train | valid | out_of_bag prediction |
|---|---|---|---|---|
| **RMSE** | **Trước EDA** | 0.41517 | 0.45419 | 0.45289 |
| | **Sau EDA** | 0.41174 | 0.45106 | 0.44888 |

Nhận xét:

- Ta thấy giá trị dự đoán RMSE trên bộ dữ liệu sau khi phân tích có giảm so với giá trị dự đoán trên bộ dữ liệu trước khi phân tích, điều này chứng tỏ chúng tôi phân tích EDA có hiệu quả.

**Neural Network:**




*Hình 10. Neural Network architecture.*

Trong thuật toán Neural Network chúng tôi cũng tiền xử lý như trong thuật toán Random Forest. Tuy nhiên, chúng tôi xử lý biến phân loại bằng phương pháp Embedding, phương pháp này đã thắng cuộc thi Rossmann sales competion trên Kaggle vào năm 2015.

Mỗi biến phân loại được đưa qua ma trận embedding trước khi đưa vào mô hình. Chọn số chiều embedding cho biến phân loại bằng cách tính min giữa 600 và 1.6*n_cat**0.56. Với n_cat là số loại trong biến phân loại.

Số chiều embedding được tạo dư 1 chiều, vì trong thư viện Fastai, các giá trị bị khuyết được điền với giá trị là Na, Na trở thành 1 loại trong biến phân loại. Trong bộ dữ liệu của chúng tôi không có giá trị bị khuyết, tuy nhiên điều này không ảnh hưởng tới kết quả của mô hình.

Số inputs đưa vào mô hình bằng tổng số chiều embedding của các biến phân loại và inputs của các biến liên tục.

Chúng tôi sử dụng 3 lớp LinBnDrop để huấn luyện mô hình, với ( inputs, outputs ) của mỗi lớp lần lượt là (inputs , 300), (300,200), (200,1). Giới hạn miền giá trị dự đoán cho biến phụ thuộc từ 0 – 100, loss function chúng tôi sử dụng là MSE và độ đo là RMSE.

Sau đó chúng tôi sử dụng hàm lr_find() trong thư viện Fastai để tìm learning rate phù hợp với mô hình, chúng tôi chọn lr = 3e-4 và phương pháp weight decay để huấn luyện mô hình.

Sau 8 epochs, giá trị RMSE chúng tôi đạt được là:

- Đối với bộ dữ liệu trước khi phân tích**: 0.354800.**
- Đối với bộ dữ liệu sau khi phân tích**: 0.319800.**

Nhận xét:

- Ta thấy giá trị RMSE trên tập valid ở bộ dữ liệu sau khi phân tích khá tốt so với giá trị RMSE trên bộ dữ liệu trước khi phân tích.
- Thư viện Fastai không đo độ chính xác trên tập, train, vì họ mặc định rằng, với mô hình đủ lớn và với thời gian huấn luyện đủ dài thì mô hình sẽ overfit trên tập train.

## 3. KẾT LUẬN

Từ dữ liệu ban đầu, tiến hành xử lý và phân loại các thuộc tính, ta thu được một clean data. Sau đó, phân tích dữ liệu bằng cách trực quan hóa các biến phân loại, thử nghiệm ANOVA và t-test trên biến liên tục để chọn ra các thuộc tính được cho là quan trọng, có ảnh hưởng tới mô hình. Việc chọn ra các thuộc tính có ảnh hưởng có vai trò rất quan trọng vì nó ảnh hưởng đến việc huấn luyện mô hình. Điển hình ta có thể thấy được qua quá trình huấn luyện trên bộ dữ liệu 2020 Yellow Taxi Trip Data bằng hai phương pháp Random Forest và Neural Network. Cả hai mô hình huấn luyện với dữ liệu chưa qua chọn lọc đều cho kết quả RMSE cao hơn so với sau khi đã phân tích. Với Random Forest cho kết quả trước và sau lần lượt là 0.45289 - 0.44888; Neural Network là 0.354800 - 0.319800.

Vì vậy, ta có thể thấy rằng việc phân tích dữ liệu rất quan trọng, giúp cho việc huấn luyện mô hình được chính xác và tối ưu hơn.

# TÀI LIỆU THAM KHẢO

**[1]** NYC Taxi & Limousine Commission. Link: https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page (3/12/2021).

**[2]** NYC OpenData. Link: https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc (3/12/2021).

**[3]** PeerJ. Link: https://peerj.com/articles/6339/ (5/12/2021)

# PHỤ LỤC PHÂN CÔNG NHIỆM VỤ

| STT | Thành viên | Nhiệm vụ |
|---|---|---|
| 1 | Nguyễn Thị Bảo Hân | Phân công nhiệm vụ từng thành viên.<br>Xử lý dữ liệu.<br>Trình bày báo cáo slide thuyết trình. |
| 2 | Văn Kim Ngân | Phân tích và trực quan dữ liệu.<br>Trình bày báo cáo word. |
| 3 | Tiêu Kim Hảo | Tìm hiểu thuật toán và huấn luyện mô hình. |