ĐỒ ÁN CUỐI KỲ: PHƯƠNG PHÁP NGHIÊN CỨU KHOA HỌC

	MSSV	Họ và tên đệm	Tên
Nhóm 25	1712919	Lê Văn	Vũ
	1712888	Nguyễn Đình	Tuyên
	1712887	Võ Nhật	Tường

MÔ HÌNH HỒI QUY LOGISTIC (LOGISTIC REGRESSION MODEL)

VÀ NHỮNG ỨNG DỤNG THỰC TẾ

MỤC LỤC:

A. MOTIVATION:

Với sự bùng nổ của công nghệ, hiện nay công nghệ thông tin đã có mặt ở hầu hết các lĩnh vực. Do đó dữ liệu ngày càng được tạo ra nhiều hơn, nhu cầu về thu thập, khai thác, phân tích,.. dữ liệu bắt đầu 'hot' dần.

- Nhiều hãng lớn đang sở hữu một nguồn dữ liệu khổng lồ, việc khai thác nguồn dữ liệu đó là cần thiết cho phát triển kinh doanh, mở rộng thị trường.
- Không những các ngành nghề kinh doanh, mà trong y tế, giáo dục, môi trường,... cũng rất cần khai thác những nguồn dữ liệu quý báu vốn có để phục vụ cho con người.

Mô hình hồi quy Logistic là một trong những mô hình rất tốt giúp ta có cái nhìn trực quan về dữ liệu và dự đoán.

Cùng xem một số ví dụ:

- ✓ Y tế: nghiên cứu về nguy cơ gãy xương đối với nam và nữ; sự liên quan giữa những người hút thuốc lá và bị ung thư phổi,...
- ✓ Giáo dục: dự đoán về phổ điểm sau mỗi kỳ thi; dự đoán số lượng hồ sơ nộp vào ngành công nghệ thông tin;...
- ✓ Kinh doanh: dự đoán giá nhà, giá vàng,...; Kinh doanh đồ uống: dự đoán mức độ ưa thích về tỷ lệ sữa và đường trong cà phê,..

B. LITERATURE REVIEW:

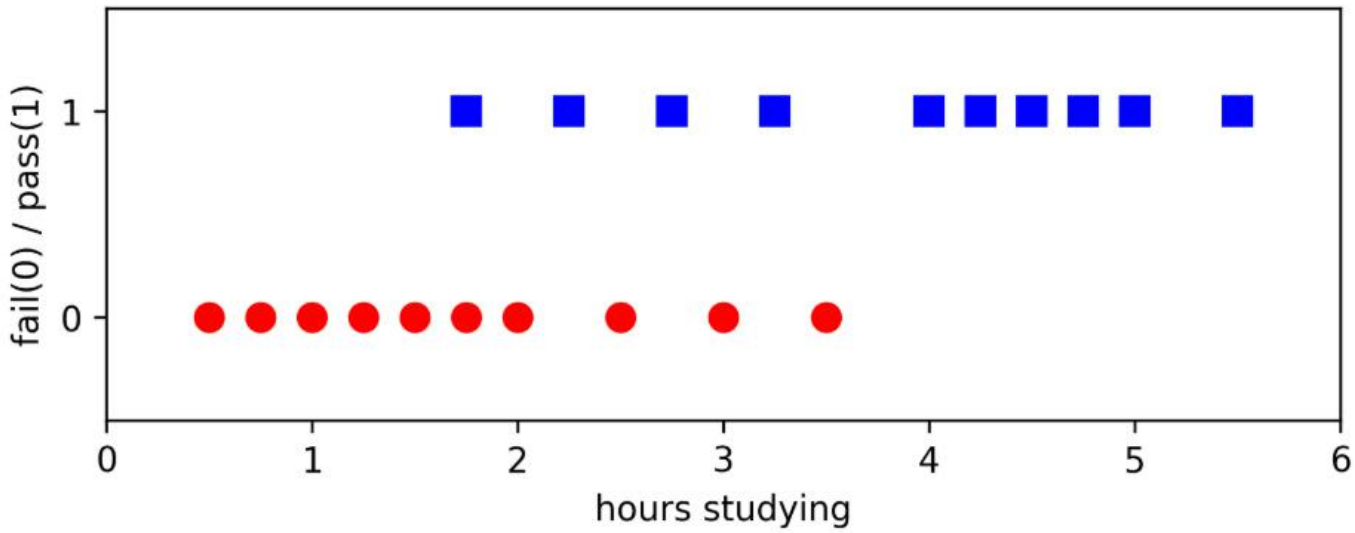
1, Vì sao nên chọn mô hình Hồi quy Logistic(Logistic Regression Model):

a) **Ví dụ dẫn nhập:** Dữ liệu về thời gian ôn thi và kết quả thi (0: rớt, 1: đỗ):

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

Nguồn: https://en.wikipedia.org/wiki/Logistic_regression

Mặc dù có một chút bất công khi học 3.5 giờ thì trượt, còn học 1.75 giờ thì lại đỗ, nhìn chung, học càng nhiều thì khả năng đỗ càng cao. PLA (Perceptron Learning Algorithm) không thể áp dụng được cho bài toán này vì không thể nói một người học bao nhiêu giờ thì 100% trượt hay đỗ, và thực tế là dữ liệu này cũng không linearly separable (điệu kiện để PLA có thể làm việc).



Hình 1: Ví dụ về kết quả thi dựa trên số giờ ôn tập.

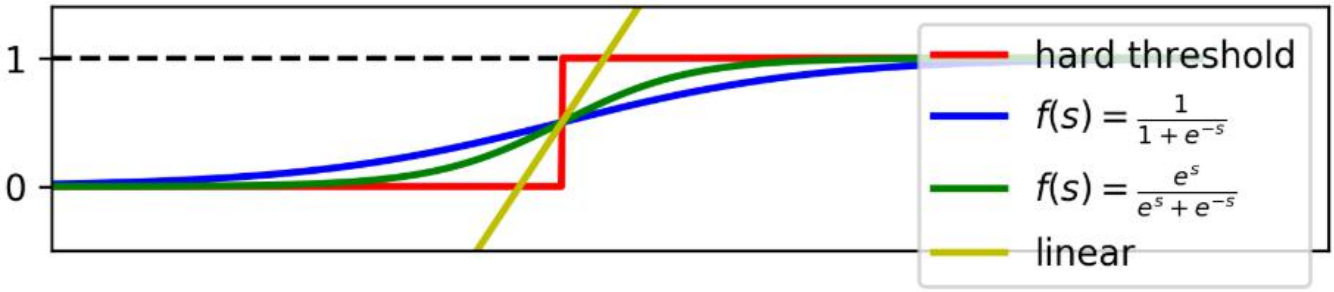
➔ Cả linear regression và PLA đều không phù hợp với bài toán này, chúng ta cần một mô hình linh hoạt hơn.

b) So sánh Linear Regression, PLA và Logistic Regression:

- Đầu ra dự đoán:

Linear Regression	$f(x) = w^T x$
PLA	$f(x) = \text{sgn}(w^T x)$
Logistic Regression	$f(x) = \theta(w^T x)$ *θ là logistic function*

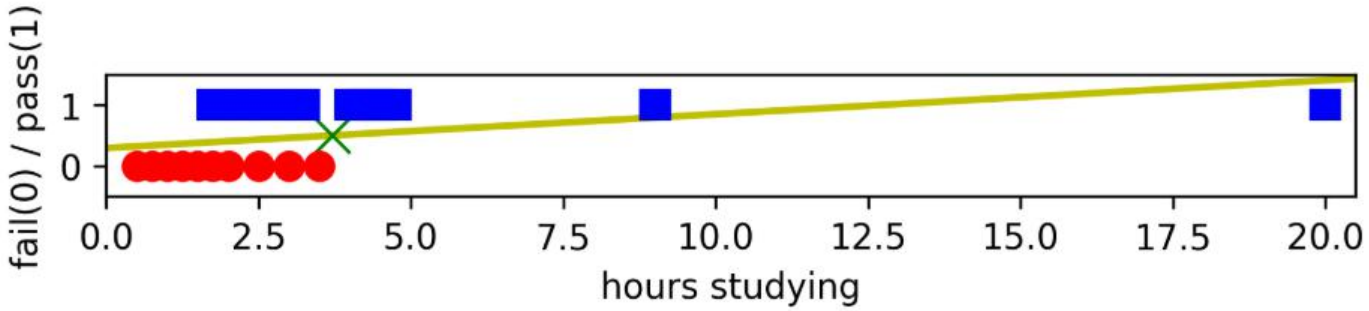
- Một số activation cho mô hình tuyến tính được cho trong hình dưới đây:



Các activation function khác nhau.

- **Đường màu vàng** biểu diễn linear regression. Đường này không bị chặn nên không phù hợp cho bài toán này. Có một *trick* nhỏ để đưa nó về dạng bị chặn: *cắt* phần nhỏ hơn 0 bằng cách

cho chúng bằng 0, *cắt* các phần lớn hơn 1 bằng cách cho chúng bằng 1. Sau đó lấy điểm trên đường thẳng này có tung độ bằng 0.5 làm điểm phân chia hai *class*, đây cũng không phải là một lựa chọn tốt. Giả sử có thêm vài bạn *sinh viên tiêu biểu* ôn tập đến 20 giờ và, tất nhiên, thi đỗ. Khi áp dụng mô hình linear regression như hình dưới đây và lấy mốc 0.5 để phân lớp, toàn bộ sinh viên thi trượt vẫn được dự đoán là trượt, nhưng rất nhiều sinh viên thi đỗ cũng được dự đoán là trượt (nếu ta coi điểm x màu xanh lục là *ngưỡng cứng* để đưa ra kết luận). Rõ ràng đây là một mô hình không tốt *(Anh chàng sinh viên tiêu biểu này đã kéo theo rất nhiều bạn khác bị trượt).*



Tại sao Linear Regression không phù hợp?

- **Đường màu đỏ** (chỉ khác với activation function của PLA ở chỗ hai class là 0 và 1 thay vì -1 và 1) cũng thuộc dạng ngưỡng cứng (hard threshold). PLA không hoạt động trong bài toán này vì dữ liệu đã cho không linearly separable.

- **Các đường màu xanh lam** và xanh lục phù hợp với bài toán của chúng ta hơn. Chúng có một vài tính chất quan trọng sau:

  ✓ Là hàm số liên tục nhận giá trị thực, bị chặn trong khoảng (0,1).
  ✓ Nếu coi điểm có tung độ là 0.5 làm điểm phân chia thì các điểm càng xa điểm này về phía bên trái có giá trị càng gần 0. Ngược lại, các điểm càng xa điểm này về phía phải có giá trị càng gần 1 ➔ khớp với nhận xét rằng học càng nhiều thì xác suất đỗ càng cao và ngược lại.
  ✓ Mượt (smooth) nên có đạo hàm mọi nơi, có thể được lợi trong việc tối ưu.

2, Mô hình Hồi quy Logistic(Logistic Regression Model):

- Gọi p là xác suất của một sự kiện, thì odd có định nghĩa: $odd = \frac{p}{1-p}$

- Mô hình hồi quy Logistic phát biểu rằng $log(odd)$ tùy thuộc vào giá trị của x thông qua một hàm số tuyến tính gồm 2 thông số sau:

$$\log(odd) = \alpha + \beta x + \varepsilon$$

$$\Leftrightarrow \log(\frac{p}{1-p}) = \alpha + \beta x + \varepsilon$$

Trong đó: log(odd) còn được gọi là *logit(p)* (do đó mới có tên là logistic); α và β là 2 thông số cần ước tính dữ liệu và ε là phần dư (residual), tức là phần không thể giải thích bằng x. Lý do hoán chuyển từ p thành *logit(p)* vì o có giá trị [0, 1], trong khi đó *logit(p)* có giá trị vô giới hạn và do đó thích hợp cho việc phân tích theo Linear Regression.

- Trong thực tế, ta không biết giá trị thật của 2 thông số α và β và phải ước tính từ số liệu quan sát được. Theo quy ước thống kê, ước số(estimates) của 2 thông số được ký hiệu hóa bằng dấu mũ: $\hat{\alpha}, \hat{\beta}$.

3, Một số ứng dụng thực tế:

3.1. Phân tích dữ liệu về độ ưa thích với hàm lượng chất béo trong nước tương:

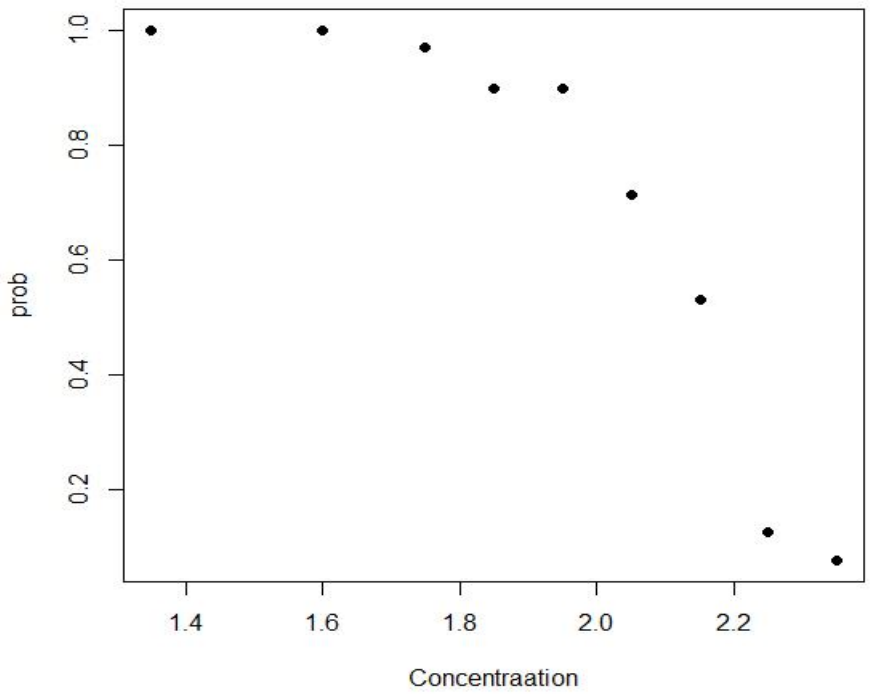
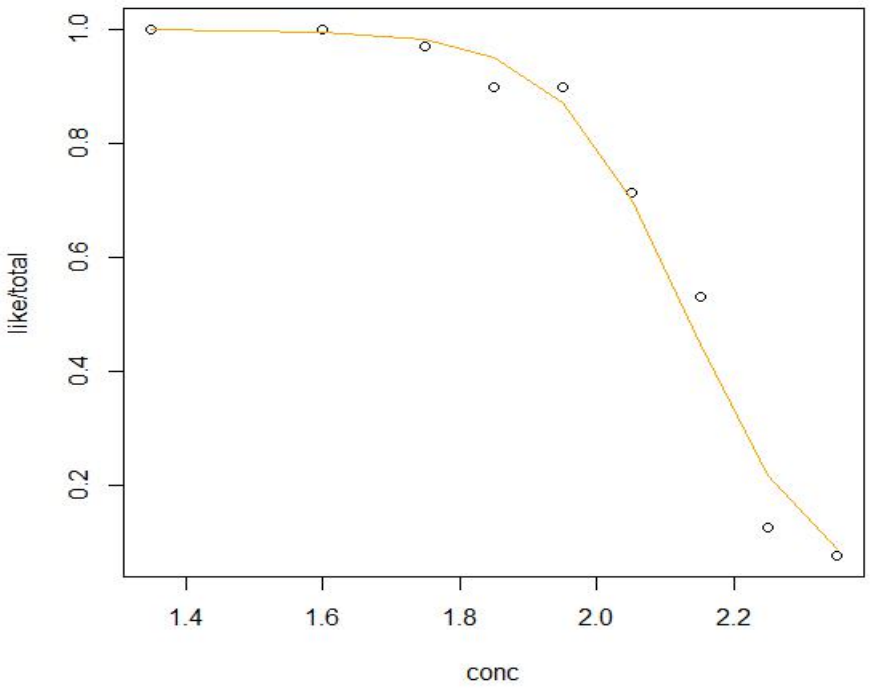
Source code đã được up lên github cá nhân: *https://github.com/vanvule/ppnckh_logistic_regression*

a, Dữ liệu mẫu:

Concentration	Like	Dislike	Total
1.35	13	0	13
1.60	19	0	19
1.75	67	2	69
1.85	45	5	50
1.95	71	8	79
2.05	50	20	70
2.15	35	31	66
2.25	7	49	56
2.35	1	12	13

Nguồn: Nguyễn Văn Tuấn, ĐH Tôn Đức Thắng.

b, Xử lý dữ liệu *(Sử dụng phần mềm R phiên bản 3.5.0)*:

Biểu đồ với các điểm là dữ liệu thực tế.	
Đường tiên lượng *(màu cam)* sau khi chạy với mô hình hồi quy Logistic rất sát với dữ liệu thực tế.	

3.2. Sự tiếp cận quảng cáo của người dùng Internet:

Source code: *https://github.com/shoaibb/Logistic-Regression*

a, Dữ liệu: Gồm 1000 mẫu với 10 trường:

- 'Daily Time Spent on Site': thời gian lướt web (phút)
- 'Age': tuổi
- 'Area Income': Thu nhập trung bình của khu vực người tiêu dùng sinh sống.
- 'Daily Internet Usage': Thời gian trung bình sử dụng Internet trên 1 ngày (phút)
- 'Ad Topic Line': Tiêu đề quảng cáo.
- 'City': Thành phố
- 'Male': Người tiêu dùng la Nam (0: không; 1: có)
- 'Country': Quốc gia
- 'Timestamp': Thời gian nhấp vào quảng cáo.
- 'Clicked on Ad': có click vào quảng cáo hay không (0: không; 1: có)

Daily Time Spent on Site	Age	Area Income	Daily Internet Usage	Ad Topic Line	City	Male	Country	Timestamp	Clicked on Ad
68.95	35	61833.9	256.09	Cloned 5thgeneration orchestration	Wrightburgh	0	Tunisia	3/27/2016 0:53	0
80.23	31	68441.85	193.77	Monitored national standardization	West Jodi	1	Nauru	4/4/2016 1:39	0
69.47	26	59785.94	236.5	Organic bottom-line service-desk	Davidton	0	San Marino	3/13/2016 20:35	0
74.15	29	54806.18	245.89	Triple-buffered reciprocal time-frame	West Terrifurt	1	Italy	1/10/2016 2:31	0
68.37	35	73889.99	225.58	Robust logistical utilization	South Manuel	0	Iceland	6/3/2016 3:36	0
59.99	23	59761.56	226.74	Sharable client-driven software	Jamieberg	1	Norway	5/19/2016 14:30	0

88.91	33	53852.85	208.36	Enhanced dedicated support	Brandonstad	0	Myanmar	1/28/2016 20:59	0
66	48	24593.33	131.76	Reactive local challenge	Port Jefferybury	1	Australia	3/7/2016 1:40	1
74.53	30	68862	221.51	Configurable coherent function	West Colin	1	Grenada	4/18/2016 9:33	0

b, Xử lý dữ liệu *(Bằng ngôn ngữ python)*:

- Chia dữ liệu thành training set và testing set:

```
In [11]:ad_data.drop(['Ad Topic Line', 'City', 'Country', 'Timestamp'], axis=1,
inplace=True)
In [12]:X = ad_data.drop(['Clicked on Ad'], axis = 1)y = ad_data['Clicked on Ad']
In [13]:from sklearn.model_selection import train_test_splitX_train, X_test,
y_train, y_test = train_test_split(X,y,test_size=0.3,random_state=101)
In [14]:from sklearn.linear_model import LogisticRegressionlogmodel =
LogisticRegression()
```

- Huấn luyện vfa phù hợp với mô hình trên trainning set:

```
In [15]:logmodel.fit(X_train, y_train)
Out[15]:LogisticRegression(C=1.0,          class_weight=None,          dual=False,
fit_intercept=True,
          intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
          penalty='l2', random_state=None, solver='liblinear', tol=0.000
1,
          verbose=0, warm_start=False)
```

- Dự đoán giá trị với test training:

```
In [16]: predictions = logmodel.predict(X_test)

In [18]: from sklearn.metrics import classification_report
In [19]: print(classification_report(y_test, predictions))
             precision    recall  f1-score   support

          0       0.91      0.95      0.93       157
          1       0.94      0.90      0.92       143

avg / total       0.92      0.92      0.92       300
```

➔ ***Ta thấy, mô hình Hồi quy Loggistic chạy rất tốt với tập dữ liệu này, với độ chính xác trên 90%***

3.3. Sự liên quan giữa GRE (Điểm thi tốt nghiệp), GPA (điểm trung bình) và uy tín của tổ chức đại học có hiệu lực nhập học vào trường sau đại học:

Source code đã được up lên github cá nhân: *https://github.com/vanvule/ppnckh_logistic_regression*

a, Dữ liệu: *nguồn: https://stats.idre.ucla.edu/stat/data/binary.csv*

Gồm 400 mẫu với 4 trường:

- 1 biến outcome nhị phân: admit.

- 3 biến còn lại là biến dự đoán: gre, gpa và rank (nhận giá trị từ 1 - 4, với 1 là mức uy tín cao nhất và 4 là thấp nhất).

Ví dụ: 6 mẫu dữ liệu đầu tiên (chạy bằng R):

```
  admit gre  gpa rank
1     0 380 3.61    3
2     1 660 3.67    3
3     1 800 4.00    1
4     1 640 3.19    4
5     0 520 2.93    4
6     1 760 3.00    2
```

b, Xử lý dữ liệu *(Sử dụng phần mềm R phiên bản 3.5.0)*:

- Kết quả khi chạy với mô hình logistic:

```
Call:
glm(formula = admit ~ gre + gpa + rank, family = "binomial",
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6268  -0.8662  -0.6388   1.1490   2.0790

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.989979   1.139951  -3.500 0.000465 ***
gre          0.002264   0.001094   2.070 0.038465 *
gpa          0.804038   0.331819   2.423 0.015388 *
rank2       -0.675443   0.316490  -2.134 0.032829 *
```

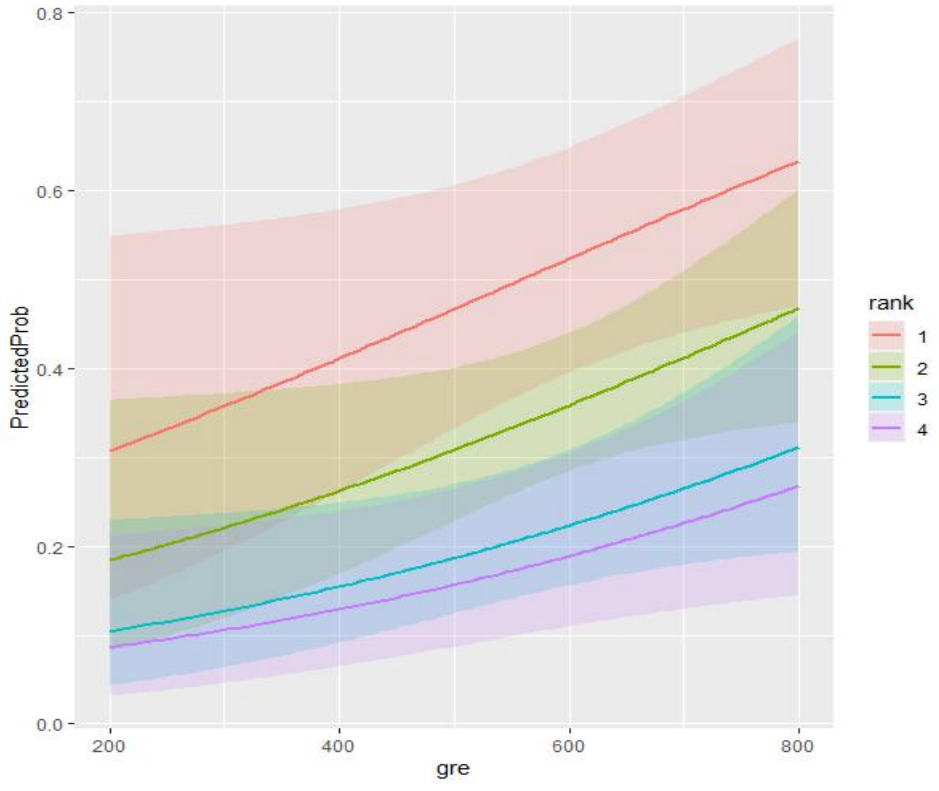
```
rank3       -1.340204   0.345306  -3.881 0.000104 ***
rank4       -1.551464   0.417832  -3.713 0.000205 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.52

Number of Fisher Scoring iterations: 4
```

- Dự đoán:

```
  gre    gpa rank     rankP
1 587.7 3.3899    1 0.5166016
2 587.7 3.3899    2 0.3522846
3 587.7 3.3899    3 0.2186120
4 587.7 3.3899    4 0.1846684
```

Trong kết quả trên, ta thấy rằng xác suất dự đoán được chấp nhận vào chương trình sau đại học là 0,52 đối với sinh viên từ các tổ chức đại học có uy tín cao nhất (rank = 1) và 0,18 cho sinh viên từ các tổ chức xếp hạng thấp nhất (rank = 4)

- Đồ thị dự đoán xác suất:



C. PROPASOL:

Ta đã biết Giả định về mô hình hồi quy Logistic như sau:

- Mô hình cung cấp một sự 'đại diện' tiêu biểu giữa biến outcome và biễn x.
- Outcomes độc lập với nhau.
- Biến tiên lượng không có sai số ngẫu nhiên.

! Nhưng trên thực tế, dữ liệu là rất đa dạng và phức tạp nên sai số là không thể tránh khỏi các sai số.

➔ Cải tiến mô hình để giải quyết vấn đề này.

✓ Hiện nay, nhiều ngôn ngữ và với những thư viện xử lý dữ liệu cho ra kết quả rất tốt):

Ví dụ: Maximum Likehood Estimator(MLE) ước tính tham số β_1,β_2 bằng cách tối đa hàm:

$$L(\beta_0,\beta_1)=\prod_{i=1}^{N}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}=\prod_{i=1}^{N}\frac{\exp(y_i(\beta_0+\beta_1x_i))}{1+\exp(\beta_0+\beta_1x_i)}$$

Có thê triển khai trong R bằng hàm **glm** và **lrm**.

D. PLAN:

STT	Nội dung	Người thực hiện	TG bắt đâu	TG kết thúc	Mức độ Hoàn thành(%)
1	**Phân tích yêu cầu**		**31/05/2020**	**25/06/202020**	
1.1	Tìm tài liệu	Vũ, Tường, Tuyên	31/05/2020	05/6/2020	100
1.2	Chọn lọc và đóng gói tài liệu	Tuyên	05/6/2020	10/6/2020	100
1.3	Tìm hiểu về mô hình hồi quy Logistic	Tường	10/6/2020	12/6/2020	100
1.4	So sánh mô hình hồi quy Logistic với các mô hình khác	Vũ	12/6/2020	15/06/2020	100
1.5	Phân tích toán học về mô hình	Vũ	15/06/2020	22/06/2020	100

1.6	Đánh giá và nhận xét về mô hình với các tập dữ liệu thực tế.	Tường, Tuyên	22/06/2020	25/06/2020	100
2	**Cài đặt**		**25/06/2020**	**20/07/2020**	
2.1	Ứng dụng số 1	Vũ	25/06/2020	30/06/2020	100
2.2	Ứng dụng số 2	Tường	30/06/2020	07/07/2020	100
2.3	Ứng dụng số 3	Vũ	07/07/2020	15/07/2020	100
2.4	Đánh giá kết quả và báo cáo	Tuyên	15/07/2020	20/07/2020	100
3	**Tổng kết**		**20/07/2020**	**01/08/2020**	
3.1	Viet báo cáo (Doc, slide)	Tuyên	20/07/2020	25/07/2020	100
3.2	Test lại tất cả các sản phẩm đã làm	Vũ (UD1+3), Tường(UD2)	25/07/2020	28/07/2020	100
3.3	Hoan thiện đồ án	Vũ, Tường, Tuyên	28/07/2020	01/08/2020	100

Nguồn tham khảo:

- Loạt bài giảng về Hồi quy Logistic của thầy Nguyễn Văn Tuấn, Đại học Tôn Đức Thắng: *https://www.youtube.com/channel/UC21dOPe-YHO3Gw6BRbyeotQ*

- *https://machinelearningcoban.com/*

- *https://github.com/shoaibb/Logistic-Regression*

- R Online Manual: *https://stat.ethz.ch/R-manual/R-patched/library/stats/html/glm.html*

References:

[1] Hosmer, D. & Lemeshow, S. (2000). Applied Logistic Regression (Second Edition). New York: John Wiley & Sons, Inc.

[2] Long, J. Scott (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

[3] Bircan H., Logistic Regression Analysis: Practice in Medical Data, Kocaeli University Social Sciences Institute Journal, 2004 / 2: 185-208

[4] Çolak, E., Özdamar K., 2004, Review of Conditional and Limited Regression Models by the Risk Factors in Fatal Traffic Accidents, OGÜ Faculty of Medicine Journal, Volume 26 P.1 Eskişehir

[5] Elhan, A.H, 1997, Review of Logistic Regression Analysis and Implementation in Medicine. (PhD thesis in biostatistics) A.U., 4-29, Ankara