TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-TPHCM
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO

Môn: Toán ứng dụng và thống kê cho công nghệ thông tin

Chủ đề: Linear Regression.

Họ tên: Kuo Yung Sheng

MSSV: 21127684

Lớp: 21CLC07

Thành phố Hồ Chí Minh, ngày 19 tháng 6 năm 2023

Mục lục

1. Thư viện sử dụng

seaborn
matplotlib.pyplot
Lý do sử dụng: để trực quan hóa dữ liệu.

sklearn.model_selection import cross_validate
Lý do sử dụng: chia tập data thành nhiều phần để tổng quát hóa.

sklearn.linear_model import LinearRegression
Lý do sử dụng: gọi LinearRegression model để huấn luyện dữ liệu.

from sklearn.metrics import mean_absolute_error
Lý do sử dụng: tính chỉ số MAE.

2. Mô tả các hàm.

LinearRegression().fit(X_trainA, y_trainA): hàm train dữ liệu từ model linearRegression.

reg.predict(X_testA): hàm dự đoán từ model đã train của sklearn.

mean_absolute_error(y_true=y_test,y_pred=y_predict): hàm tính MAE.

cross_validate(LinearRegression(), X_train[[i]], y_train, cv=5,scoring='neg_mean_absolute_error'): hàm chia k-fold và tính MAE dựa vào mô hình và data truyền vào.

3. Kết quả

1a)

Huấn luyện 11 features một lần duy nhất và predict trên tập test.

```
features=['Gender', '10percentage', '12percentage', 'CollegeTier', 'Degree',
 'collegeGPA', 'CollegeCityTier', 'English', 'Logical', 'Quant', 'Domain']
X_trainA=X_train[features]
X_testA=X_test[features]
y_trainA=y_train
reg = LinearRegression().fit(X_trainA, y_trainA)
y_predict=reg.predict(X_testA)
```

Ta đính được chỉ số MAE.

```
mae = round(mean_absolute_error(y_true=y_test,y_pred=y_predict),3)
print(mae)
```

```
105052.53
```

1b)

STT	Mô hình với 1 đặc trưng	MAE
1	conscientiousness	124444.487
2	agreeableness	123813.287
3	extraversion	123914.505
4	nueroticism	123738.525
5	openess_to_experience	124119.481

Nhận xét:

Các mô hình đặc trưng tính cách chênh lệch không nhiều trong khoản 1000.

Thấp nhất là nueroticism và cao nhất là conscientiousness.

Từ chỉ số MAE ta dự đoán mô hình với đặc trưng nueroticism tốt nhất vì MAE thấp nhất.

Train lại model và tính chỉ số MAE trên tập test.

```
reg = LinearRegression().fit(X_train[['nueroticism']], y_train)
y_predict=reg.predict(X_test[['nueroticism']])
```

```
# Gọi hàm MAE (tự cài đặt hoặc từ thư viện) trên tập kiểm tra với m
mae = round(mean_absolute_error(y_true=y_test,y_pred=y_predict),3)
print(mae)
```

```
119361.917
```

1c)

STT	Mô hình với 1 đặc trưng	MAE
1	English	120963.069
2	Logical	120037.719
3	Quant	117461.464

Nhận xét

English và Logical có chỉ số MAE gần bằng nhau. Riêng Quant thì có chỉ số thấp nhất.

Từ chỉ số MAE ta dự đoán mô hình với đặc trưng Quant tốt nhất vì MAE thấp nhất.

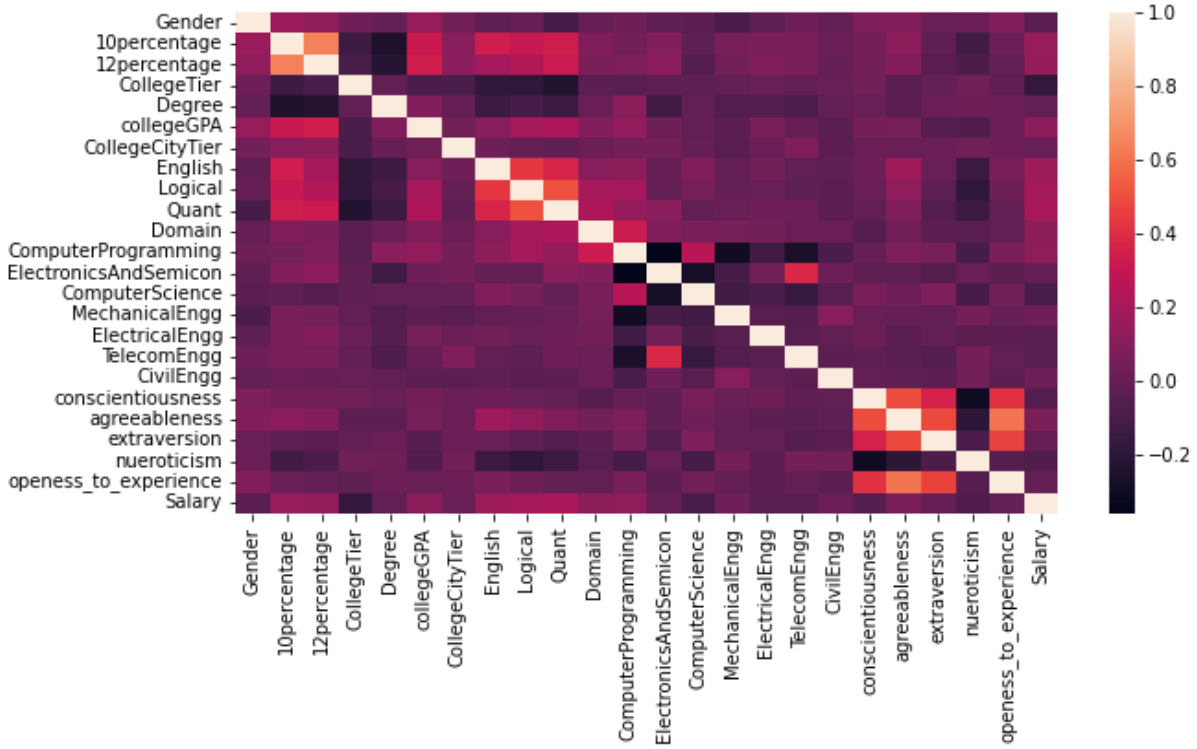
Train lại model và tính chỉ số MAE trên tập test.

```
reg = LinearRegression().fit(X_train[['Quant']], y_train)
y_predict=reg.predict(X_test[['Quant']])
```

```
# Gọi hàm MAE (tự cài đặt hoặc từ thư viện) trên tập kiểm tra với mô hình best_skill_feature_model
mae = round(mean_absolute_error(y_true=y_test,y_pred=y_predict),3)
print(mae)
```

```
108814.06
```

1d)



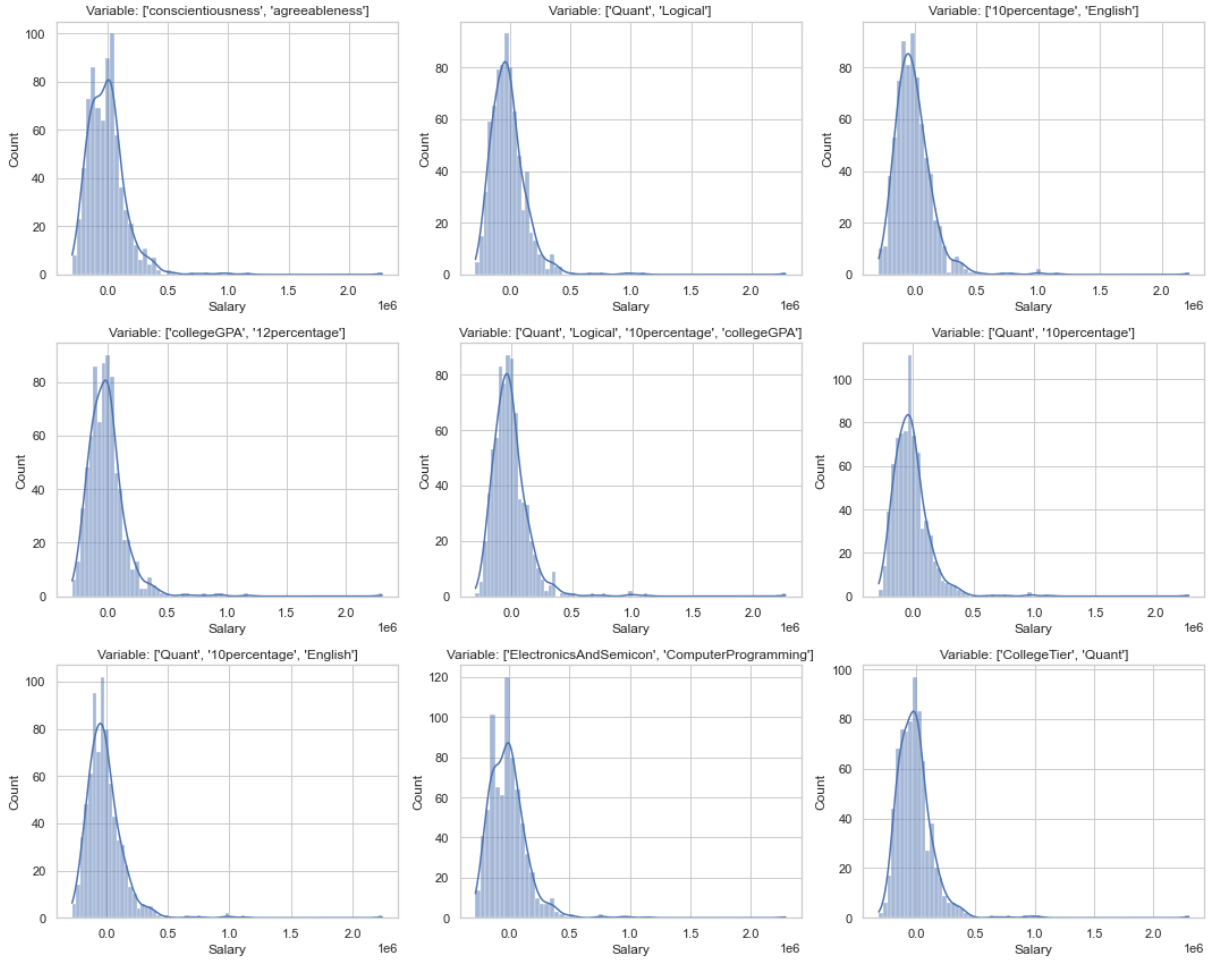
Từ bảng heatmap cho ta thấy sự tương quan giữa các đặc trưng.

Ta có thể thấy một số đặc trưng nổi trội là: '10percentage', 'CollegeTier', 'collegeGPA', 'English', 'Logical', 'Quant', 'Domain'.

Chọn ngẫu nhiên các đặc trưng này để xem kết quả cross-validate

```
['conscientiousness', 'agreeableness']
123706.57628368158
['Quant', 'Logical']
116718.94381198932
['10percentage', 'English']
117921.55684703798
['collegeGPA', '12percentage']
119178.21668003441
['Quant', 'Logical', '10percentage', 'collegeGPA']
114560.24465161569
['Quant', '10percentage']
115138.80817310812
['Quant', '10percentage', 'English']
114760.92647901442
['ElectronicsAndSemicon', 'ComputerProgramming']
122410.64303667788
['CollegeTier', 'Quant']
117475.62879273777
```

Để xác nhận lại thêm lần nữa ta vẽ biểu đồ cột thể hiện độ lệch của từng mô hình.



Ta nhận thấy rằng mô hình ['Quant','10percentage'] có kết quả tốt nhất. Huấn luyện lại mô hình trên toàn bộ tập huấn luyện.

```
# Huấn luyện lại mô hình my_best_model trên toàn bộ tập huấn luyện
reg = LinearRegression().fit(X_train[['Quant','10percentage']], y_train)
y_predict=reg.predict(X_test[['Quant','10percentage']])
  mae = round(mean_absolute_error(y_true=y_test,y_pred=y_predict),3)
  print(mae)
```

```
106022.179
```

4. Tài liệu tham khảo

https://www.kaggle.com/code/sonawanelalitsunil/graduate-salary-prediction-xgboost#visualizing-Difference

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html

https://scikit-learn.org/stable/modules/cross_validation.html