Course: Data Analysis

Control Work №2

Student’s Name and Surname **Shmakov Maxim**

Please do the tasks below.

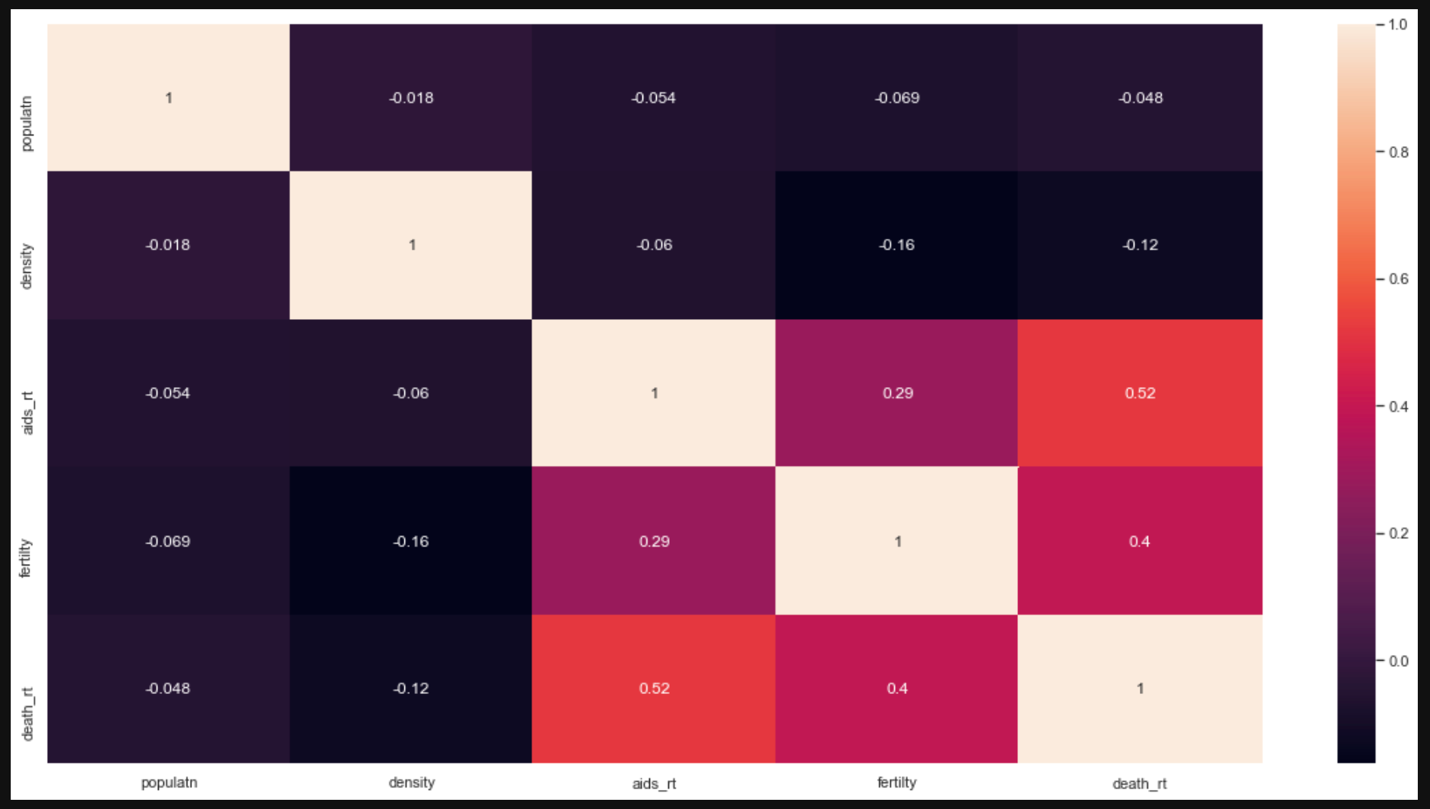
**Task 1 (8 points)**

Open **World\_data.dta** file. Select not less than 5 variables for the cluster analysis.

Selected variables: "populatn", "density", "aids\_rt", "fertilty", "death\_rt"

1.1. Why do you think that the selected variables can be used in cluster analysis (1 point)?

Those variables were selected because they have a small correlation rate between them, they all are numeric and there are no missing values in them. Correlation heatmap:

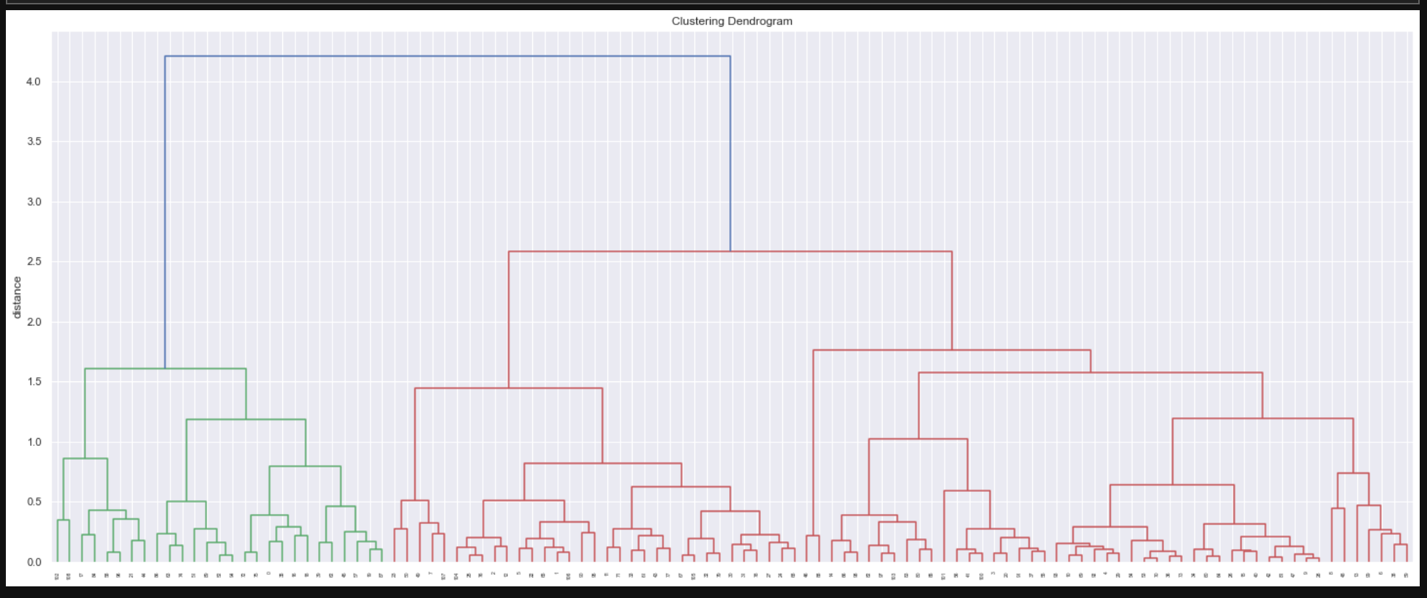


1.2. Select an appropriate clustering method and explain your selection (1 point).

Firstly, we need to construct the hierarchical clustering dendrogram to visualize the number if clusters in our data. Then the appropriate clustering method in our case would be kMeans because it can select distinct non-overlapping clusters where each data point belongs to only one group and it is what we need.

1.3. Define the number of clusters and explain your decision (2 points).

Hierarchical clustering dendrogram:



From this diagram we can see three distinct clusters, let’s construct the clusters using kMeans.

1.4. Describe each cluster in details. *It’s not enough to copy the table with descriptive statistics* (2 points).

For the first cluster we have:



For the second cluster we have:



For the third cluster we have:

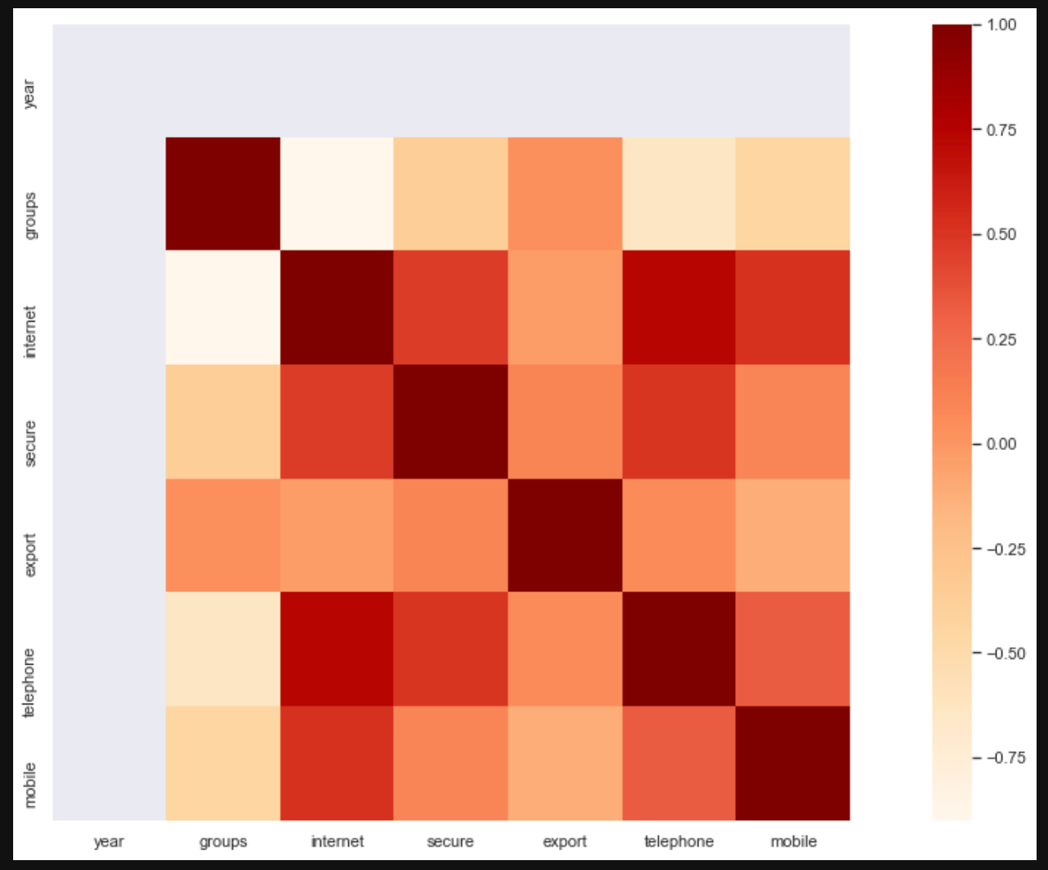


1.5. Give names to your clusters based on the description (2 point).

The names can be constructed from the level of country: countries of first world, second world and third world.

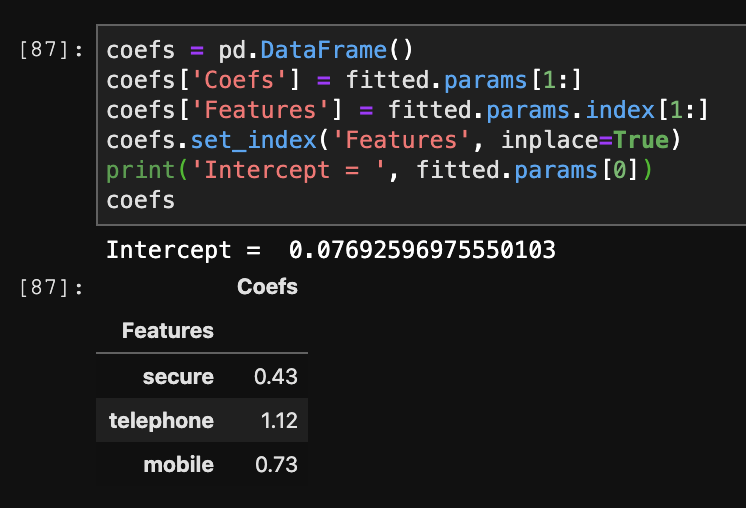
**Task 2 (10 points)**

Open **Worldbank\_data.dta** file. Do the multiple linear regression analysis using the following variables: internet, secure, export, import, telephone, high-tech, mobile. Select one dependent variable and 3 independent variables.



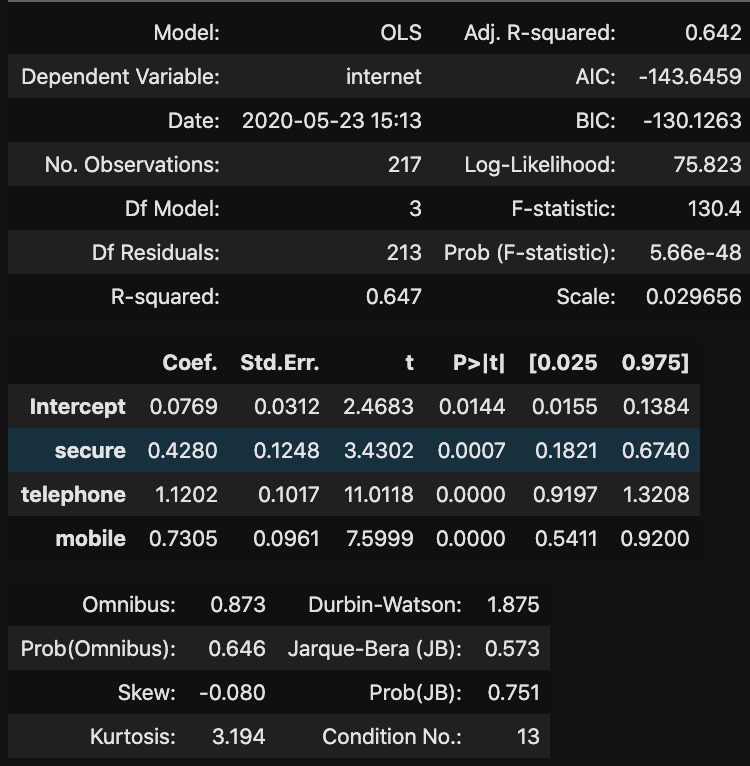
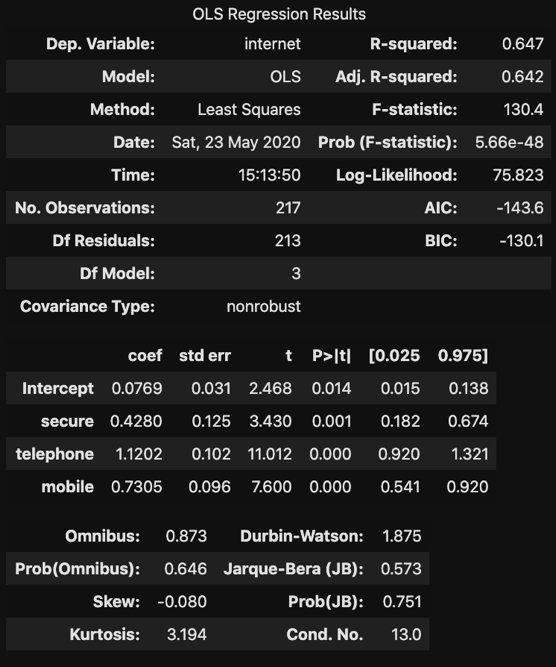
As a dependent variable I choose **internet** as it has the highest correlation rates (from heatmap). As for independent variables, I choose **secure**, **telephone** and **mobile** as they have the highest correlation with the **internet**.

2.1. Write down the regression equation for you model (1 point).



internet = 0.076926 + 0.428\*secure + 1.120\*telephone + 0.731\*mobile

2.2. Indicate the percent of the variance of the dependent variable that is explained by the model (1 point)?



~64% (adjusted R^2 = 0.642).

2.3. Which regression coefficients are statistically significant (1 point)?

All (P is less than 0.05)

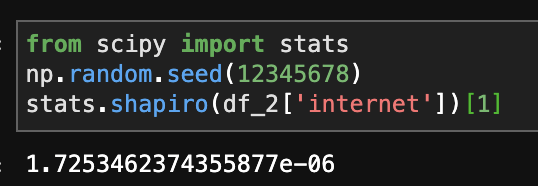
2.4. Which of the independent variables has the highest correlation with the dependent variable (1 point)?

**telephone** has the highest correlation (from heatmap 0.75).

2.5. Interpret the influence of any predictor variable on the dependent variable. *Not only the direction, but also the specificity of the relationship* (2 points).

2.6. Do the diagnostics of the model (4 points).

2.6.1. Are the residuals normally distributed (1 point)?



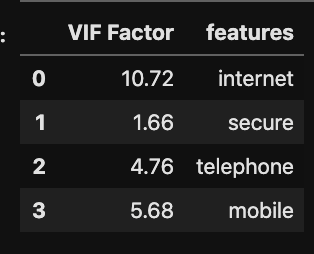
P-value is very small, so they are normally distributed.

2.6.2. Are there any outliers or influential cases? If yes, how many (1 point)?

There are three outliers: Lichteshtein, Monaco, Bermuda.

2.6.3. Check the multicollinearity and heteroscedasticity (2 points).

There is no multicollinearity between independent variables due to the fact that VIF < 10, so that is good.

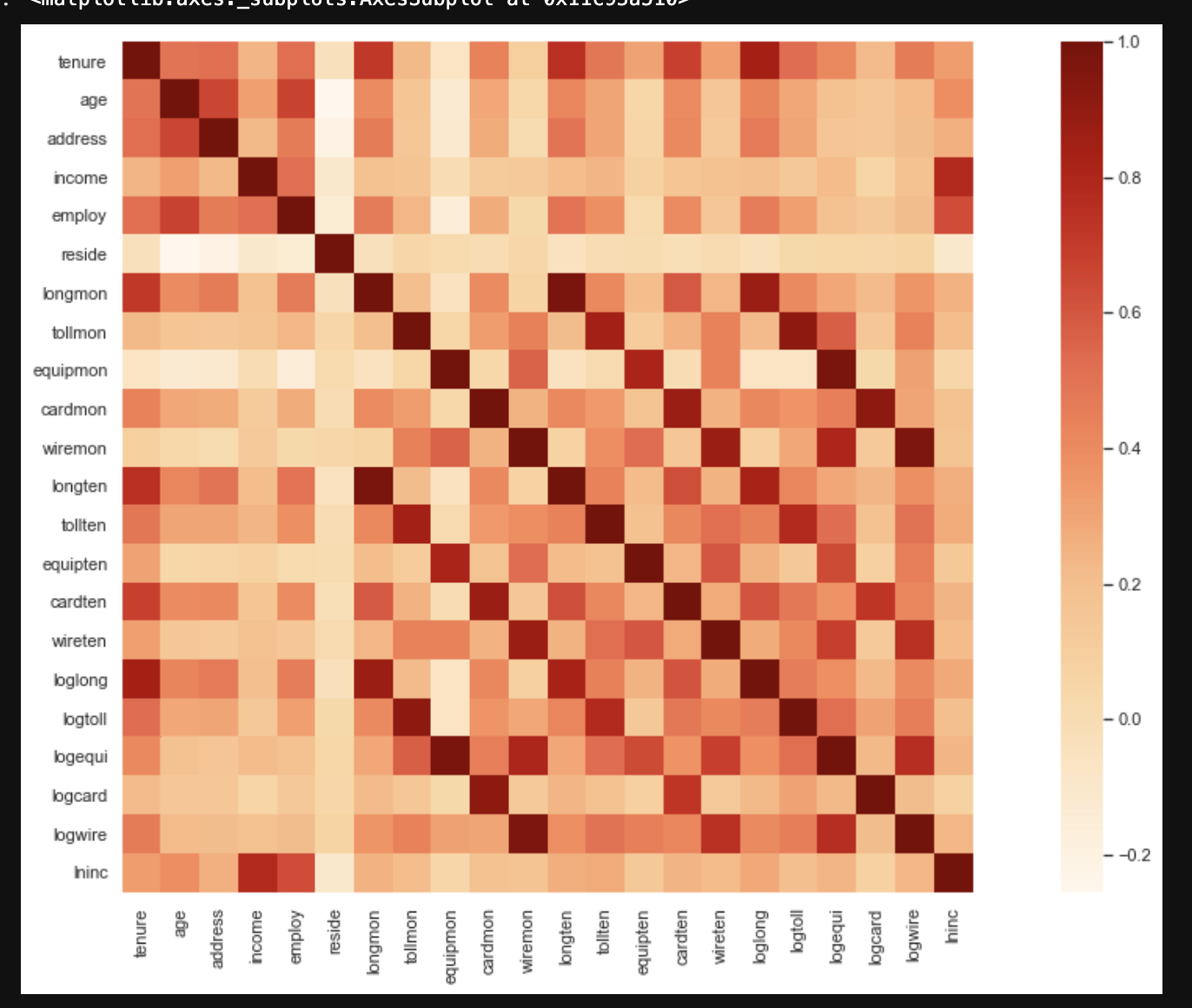


There is also no heteroscedasticity due to Breusch-Pagan / Cook-Weisberg test (P < 0.05)

**Task 3 (10 points)**

Open **telco.dta** file. Select one dependent variable. You can choose any of the following variables:

* Multiple lines
* Voice mail
* Paging service
* Internet
* Caller ID
* Call waiting
* Call forwarding
* 3-way calling
* Electronic billing



Same reasons:

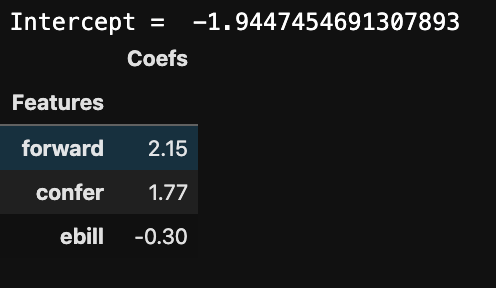
Dependent: callwait

Independent: forward confer ebill

Choose any predictor variables (not less than three).

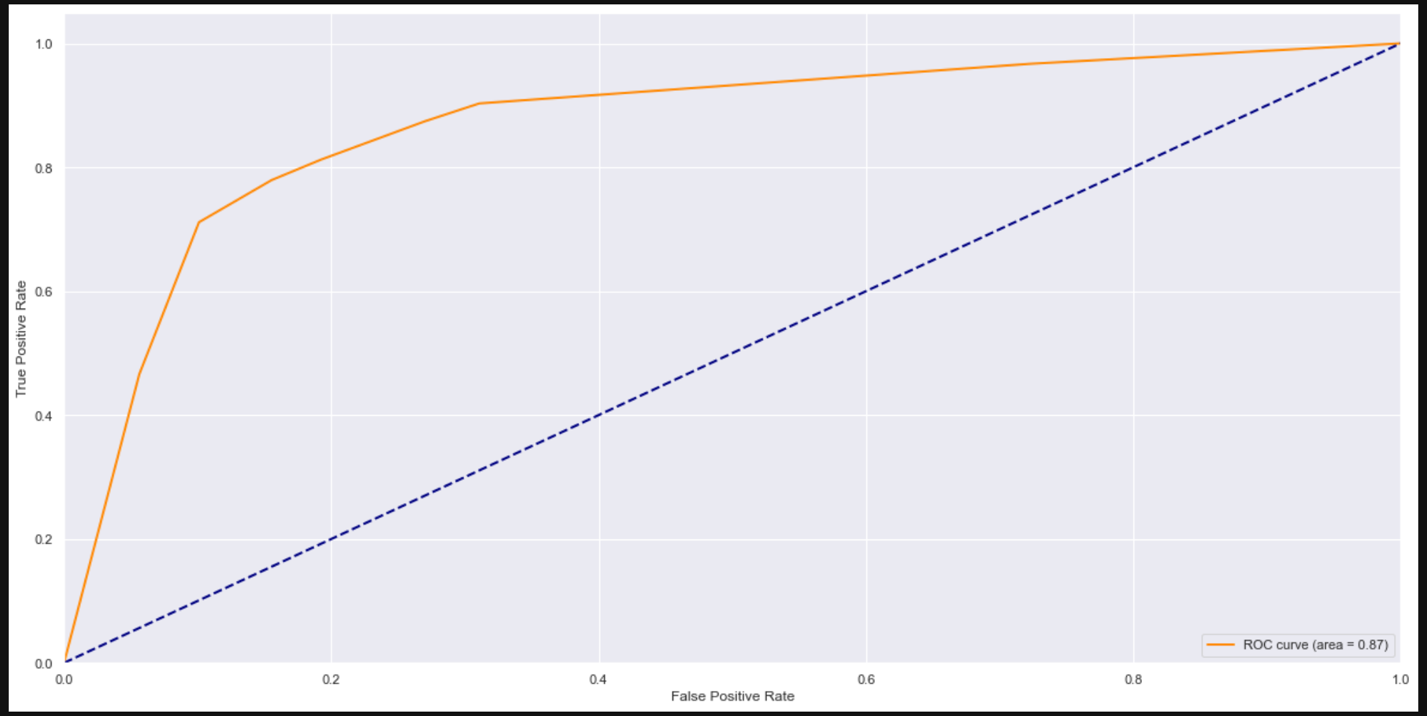
Use the binary logistic regression to analyze the data. Interpret the characteristics of the selected model.

3.1. Write down the regression equation (the linear part of the formula) (1 point).



callwait = -1.944745 + 2.153\*forward + 1.769\*confer + -0.300\*ebill

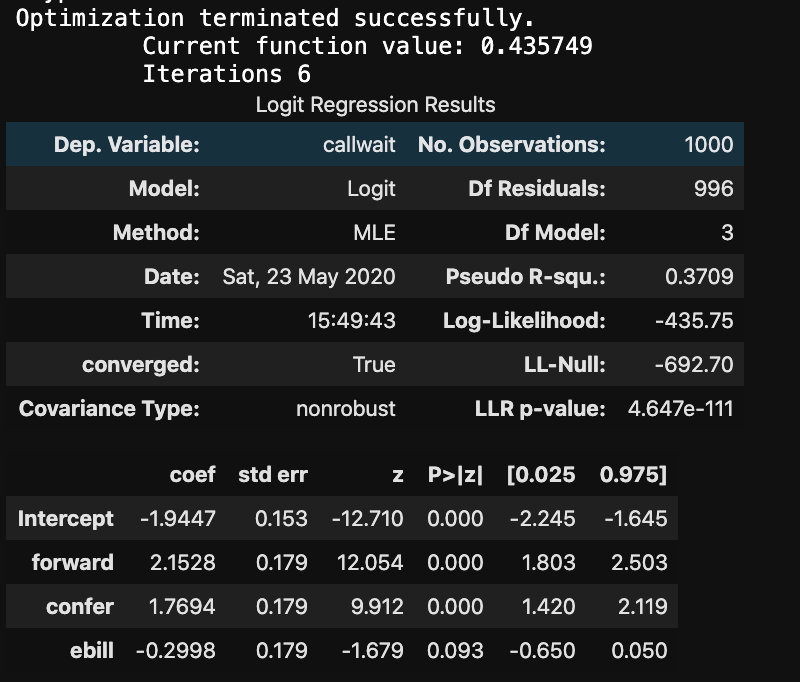
3.2. Assess the goodness-of-fit of the model and interpret the results of the analysis (2 points).



3.3. Interpret the influence of any predictor variable on the dependent variable using Exp(b). *Not only the direction, but also the specificity of the relationship* (1 point).

Answer

3.4. Which coefficients of the model are statistically significant and why (2 points)?



Ebill as it is P > 0.05 ()

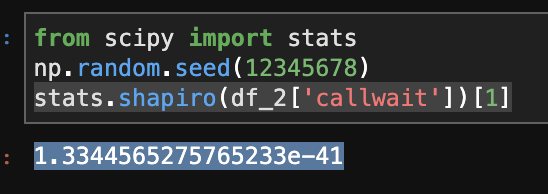
3.5. What is the percentage of correctly predicted cases by the model (1 point)?

Answer

3.6. Do the diagnostics of the model (3 points).

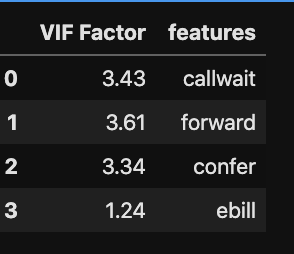
3.6.1. Are the residuals normally distributed (1 point)?

P is 1.3344565275765233e-41, so normally, yes.



3.6.2. Does the model have outliers or influential cases? If yes, how many? (1 point)

3.6.3. Test the multicollinearity (1 point).



No multicollinearity (all VIFs < 10)

|  |  |
| --- | --- |
| Total points (out of 28) | Final grade (out of 10) |
| 27 – 28 | 10 |
| 24 – 26 | 9 |
| 21 – 23 | 8 |
| 18 – 20 | 7 |
| 15 – 17 | 6 |
| 12 – 14 | 5 |
| 9 – 11 | 4 |
| 6 – 8 | 3 |
| 3 – 5 | 2 |
| 0 – 2 | 1 |