**Part I: Research Question**

Research Question: *Which variables are significant in predicting whether a patient is readmitted or not?*

The objectives of this data analysis are to learn how we can predict the probability of whether a patient will be readmitted. Thus, if we know which patients are at higher risk of readmission, we can adjust our care accordingly and plan as such. This will allow us to be more prepared, avoid penalties for readmissions, and hopefully, provide better overall healthcare to patients.

**Part II: Method Justification**

A logistic regression model assumes several stipulations. First, is based on the assumption that the dependent variable is binary (i.e. 2 different outcomes). Furthermore, the model only predicts the probability of each outcome, rather than the outcome itself. Like a multiple linear regression model, it also assumes that no multicollinearity exists among the independent variables, and that no extreme outliers in the dataset are influencing the model.

I’ll be using Python programming language; specifically, I’ll employ and utilize the pandas, NumPy, Matplotlib, Seaborn, and sklearn libraries. I’m familiar with Python and prefer to use it in a Jupyter notebook. As a general-purpose language, I know I will be able to do what I need to do. The code and syntax in Python and pandas is readable and comprehensible. Sklearn will allow me to use functions to build a regression model, while Seaborn and Matplotlib will allow visualization techniques that will be helpful for further understanding the data. Most importantly, I have built regression models using Python and the rest of these libraries/packages before, so I will be comfortable doing so again.

Logistic regression is an appropriate technique to analyze the research question summarized in Part I because our dependent variable, ReAdmis, is binary, with two possible outcomes. Therefore, we will be able to predict the probability of the outcome being yes/no using multiple independent variables.

**Part III: Data Preparation**

My data preparation goals begin with confirming the data is complete and accurate (i.e. clean), and if not, cleaning it as necessary. I will reduce the data to only the variables that I plan to use for my regression model. At that point, I will want to check for any outliers that will excessively influence my model and remove the row that holds the outlier. Since I know I’ll be working with several categorical values, I will need to convert all of those to dummy variables so they will work with the model. This will create several additional columns - I will also need to remove the original converted column. These manipulations will allow for data that is conducive to creating a multiple regression model.

Summary statistics for all variables are below. The blue highlighted variable is our response (dependent) variable, and the others are our predictor (independent) variables. The first table shows continuous variables, while the following table (on the next page) show categorical variables.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Lat** | 10000 | 38.75 | 5.40 | 17.97 | 35.26 | 39.42 | 42.04 | 70.56 |
| **Lng** | 10000 | -91.24 | 15.21 | -174.21 | -97.35 | -88.40 | -80.44 | -65.29 |
| **Population** | 10000 | 9965.25 | 14824.76 | 0 | 694.75 | 2769 | 13945 | 122814 |
| **Children** | 10000 | 2.10 | 2.16 | 0 | 0 | 1 | 3 | 10 |
| **Age** | 10000 | 53.51 | 20.64 | 18 | 36 | 53 | 71 | 89 |
| **Income** | 10000 | 40490.50 | 28521.15 | 154.08 | 19598.78 | 33768.42 | 54296.4 | 207249 |
| **VitD\_levels** | 10000 | 17.96 | 2.02 | 9.81 | 16.63 | 17.95 | 19.35 | 26.39 |
| **Doc\_visits** | 10000 | 5.01 | 1.05 | 1 | 4 | 5 | 6 | 9 |
| **Full\_meals\_eaten** | 10000 | 1.00 | 1.01 | 0 | 0 | 1 | 2 | 7 |
| **vitD\_supp** | 10000 | 0.40 | 0.63 | 0 | 0 | 0 | 1 | 5 |
| **Initial\_days** | 10000 | 34.46 | 26.31 | 1.00 | 7.90 | 35.84 | 61.16 | 71.98 |
| **Additional\_charges** | 10000 | 12934.53 | 6542.60 | 3125.7 | 7986.488 | 11573.98 | 15626.49 | 30566.1 |
| **Timely\_admission** | 10000 | 3.52 | 1.03 | 1 | 3 | 4 | 4 | 8 |
| **Timely\_treatment** | 10000 | 3.51 | 1.03 | 1 | 3 | 3 | 4 | 7 |
| **Timely\_visits** | 10000 | 3.51 | 1.03 | 1 | 3 | 4 | 4 | 8 |
| **Reliability** | 10000 | 3.52 | 1.04 | 1 | 3 | 4 | 4 | 7 |
| **Options** | 10000 | 3.50 | 1.03 | 1 | 3 | 3 | 4 | 7 |
| **Hours\_of\_treatment** | 10000 | 3.52 | 1.03 | 1 | 3 | 4 | 4 | 7 |
| **Courteous\_staff** | 10000 | 3.49 | 1.02 | 1 | 3 | 3 | 4 | 7 |
| **Active\_listening** | 10000 | 3.51 | 1.04 | 1 | 3 | 3 | 4 | 7 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **count** | **unique** | **top** | **freq** |
| **Area** | 10000 | 3 | Rural | 3369 |
| **Marital** | 10000 | 5 | Widowed | 2045 |
| **Gender** | 10000 | 3 | Female | 5018 |
| **Soft\_drink** | 10000 | 2 | No | 7425 |
| **Initial\_admin** | 10000 | 3 | Emergency Admission | 5060 |
| **HighBlood** | 10000 | 2 | No | 5910 |
| **Stroke** | 10000 | 2 | No | 8007 |
| **Overweight** | 10000 | 2 | Yes | 7094 |
| **Arthritis** | 10000 | 2 | No | 6426 |
| **Diabetes** | 10000 | 2 | No | 7262 |
| **Hyperlipidemia** | 10000 | 2 | No | 6628 |
| **Anxiety** | 10000 | 2 | No | 6785 |
| **Allergic\_rhinitis** | 10000 | 2 | No | 6059 |
| **Reflux\_esophagitis** | 10000 | 2 | No | 5865 |
| **Asthma** | 10000 | 2 | No | 7107 |
| **Services** | 10000 | 4 | Blood Work | 5265 |
| **Complication\_risk** | 10000 | 3 | Medium | 4517 |
| **BackPain** | 10000 | 2 | No | 5886 |
| **ReAdmis** | 10000 | 2 | No | 6331 |

See attached Jupyter notebook for data preparation steps and annotated code.

**Univariate Exploration:**

Chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Graphical user interface, application

Description automatically generated

A picture containing text, crossword puzzle, scoreboard, vector graphics

Description automatically generated

**Bivariate Exploration:**

A picture containing scatter chart

Description automatically generated

A picture containing shape

Description automatically generated Shape, arrow

Description automatically generated

A picture containing icon

Description automatically generated Shape

Description automatically generated with medium confidence

Diagram, icon

Description automatically generated with medium confidence A picture containing chart

Description automatically generated

Shape

Description automatically generated Diagram, icon

Description automatically generated with medium confidence

A picture containing text, crossword puzzle

Description automatically generated

A picture containing text, crossword puzzle

Description automatically generated

**Part IV: Model Comparison and Analysis**

**Initial Model:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model:** | Logit | **Pseudo R-squared:** | inf |  |  |  |
| **Dependent Variable:** | ReAdmis Yes | **AIC:** | inf |  |  |  |
| **Date:** | 9/19/21 15:14 | **BIC:** | inf |  |  |  |
| **No. Observations:** | 10000 | **Log-Likelihood:** | -inf |  |  |  |
| **Df Model:** | 47 | **LL-Null:** | 0 |  |  |  |
| **Df Residuals:** | 9952 | **LLR p-value:** | 1 |  |  |  |
| **Converged:** | 1 | **Scale:** | 1 |  |  |  |
| **No. Iterations:** | 14 |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | **Coef.** | **Std.Err.** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **intercept** | -85.5937 | 5.1785 | -16.5287 | 0 | -95.7433 | -75.4441 |
| **Initial\_days** | 1.5392 | 0.0891 | 17.2792 | 0 | 1.3646 | 1.7138 |
| **Lat** | 0.0405 | 0.0199 | 2.0386 | 0.0415 | 0.0016 | 0.0794 |
| **Lng** | 0.0089 | 0.0069 | 1.2848 | 0.1989 | -0.0047 | 0.0225 |
| **Population** | 0 | 0 | 0.9183 | 0.3584 | 0 | 0 |
| **Children** | 0.0917 | 0.0476 | 1.9292 | 0.0537 | -0.0015 | 0.185 |
| **Age** | -0.0044 | 0.0153 | -0.2875 | 0.7737 | -0.0344 | 0.0256 |
| **Income** | 0 | 0 | 0.4294 | 0.6676 | 0 | 0 |
| **Doc\_visits** | 0.0126 | 0.1016 | 0.1239 | 0.9014 | -0.1866 | 0.2118 |
| **Full\_meals\_eaten** | 0.0339 | 0.1067 | 0.3178 | 0.7506 | -0.1753 | 0.2431 |
| **vitD\_supp** | -0.0985 | 0.1695 | -0.5813 | 0.561 | -0.4306 | 0.2336 |
| **Timely\_admission** | -0.0155 | 0.1568 | -0.0987 | 0.9214 | -0.3229 | 0.2919 |
| **VitD\_levels** | 0.0313 | 0.0503 | 0.6218 | 0.5341 | -0.0674 | 0.13 |
| **Additional\_charges** | 0 | 0.0001 | 0.4867 | 0.6265 | -0.0001 | 0.0002 |
| **Timely\_treatment** | 0.2697 | 0.1407 | 1.917 | 0.0552 | -0.006 | 0.5455 |
| **Timely\_visits** | -0.1553 | 0.1308 | -1.1873 | 0.2351 | -0.4116 | 0.1011 |
| **Reliability** | 0.0529 | 0.1159 | 0.4564 | 0.6481 | -0.1743 | 0.2801 |
| **Options** | -0.1545 | 0.1205 | -1.2819 | 0.1999 | -0.3908 | 0.0817 |
| **Hours\_of\_treatment** | -0.0687 | 0.1283 | -0.5354 | 0.5923 | -0.32 | 0.1827 |
| **Courteous\_staff** | 0.1279 | 0.1166 | 1.0967 | 0.2728 | -0.1007 | 0.3566 |
| **Active\_listening** | -0.2143 | 0.1099 | -1.9494 | 0.0513 | -0.4298 | 0.0012 |
| **Area\_Suburban** | 0.0525 | 0.258 | 0.2034 | 0.8389 | -0.4532 | 0.5581 |
| **Area\_Urban** | 0.0818 | 0.2658 | 0.3077 | 0.7583 | -0.4392 | 0.6028 |
| **Marital\_Married** | 0.1327 | 0.3425 | 0.3875 | 0.6984 | -0.5386 | 0.8041 |
| **Marital\_Never\_Married** | 0.2749 | 0.35 | 0.7854 | 0.4322 | -0.4111 | 0.961 |
| **Marital\_Separated** | -0.0991 | 0.3526 | -0.281 | 0.7787 | -0.7901 | 0.5919 |
| **Marital\_Widowed** | 0.0952 | 0.3449 | 0.2759 | 0.7826 | -0.5809 | 0.7712 |
| **Gender\_Male** | 0.1766 | 0.2142 | 0.8244 | 0.4097 | -0.2432 | 0.5964 |
| **Gender\_Nonbinary** | 0.2004 | 0.7361 | 0.2722 | 0.7855 | -1.2425 | 1.6432 |
| **Soft\_drink\_Yes** | 0.2667 | 0.2486 | 1.0731 | 0.2832 | -0.2205 | 0.7539 |
| **Initial\_admin\_Emergency Admission** | 2.6205 | 0.2987 | 8.7721 | 0 | 2.035 | 3.2061 |
| **Initial\_admin\_Observation Admission** | 0.8808 | 0.2927 | 3.0093 | 0.0026 | 0.3072 | 1.4545 |
| **HighBlood\_Yes** | 0.5593 | 0.5992 | 0.9335 | 0.3506 | -0.6151 | 1.7337 |
| **Stroke\_Yes** | 1.7501 | 0.2809 | 6.2297 | 0 | 1.1995 | 2.3008 |
| **Overweight\_Yes** | -0.2793 | 0.2354 | -1.1869 | 0.2353 | -0.7407 | 0.182 |
| **Arthritis\_Yes** | -1.4067 | 0.2346 | -5.9961 | 0 | -1.8665 | -0.9469 |
| **Diabetes\_Yes** | 0.51 | 0.2387 | 2.1366 | 0.0326 | 0.0421 | 0.9778 |
| **Hyperlipidemia\_Yes** | 0.2919 | 0.221 | 1.3212 | 0.1864 | -0.1412 | 0.725 |
| **Anxiety\_Yes** | -1.0589 | 0.2324 | -4.5556 | 0 | -1.5145 | -0.6033 |
| **Allergic\_rhinitis\_Yes** | -0.4201 | 0.2174 | -1.9321 | 0.0533 | -0.8462 | 0.0061 |
| **Reflux\_esophagitis\_Yes** | -0.3484 | 0.2192 | -1.5894 | 0.112 | -0.7781 | 0.0812 |
| **Asthma\_Yes** | -1.4374 | 0.2437 | -5.8991 | 0 | -1.915 | -0.9598 |
| **Services\_CT\_Scan** | 1.5669 | 0.3751 | 4.1772 | 0 | 0.8317 | 2.3021 |
| **Services\_Intravenous** | -0.0223 | 0.2372 | -0.094 | 0.9251 | -0.4871 | 0.4426 |
| **Services\_MRI** | 2.7175 | 0.5131 | 5.2967 | 0 | 1.7119 | 3.7231 |
| **Complication\_risk\_Low** | -1.9305 | 0.307 | -6.2886 | 0 | -2.5322 | -1.3288 |
| **Complication\_risk\_Medium** | -0.3527 | 0.2441 | -1.4448 | 0.1485 | -0.8311 | 0.1257 |
| **BackPain\_Yes** | 0.2664 | 0.214 | 1.2447 | 0.2132 | -0.1531 | 0.6858 |

Our model above uses a lot of variables, and several of them have p-values above 0.05. We should take all of those out to create a cleaner model.

**Reduced Model:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model: | Logit | Pseudo R-squared: | 0.947 |  |  |  |
| Dependent Variable: | ReAdmis Yes | AIC: | 724.122 |  |  |  |
| Date: | 9/20/21 10:24 | BIC: | 817.8564 |  |  |  |
| No. Observations: | 10000 | Log-Likelihood: | -349.06 |  |  |  |
| Df Model: | 12 | LL-Null: | -6572.9 |  |  |  |
| Df Residuals: | 9987 | LLR p-value: | 0 |  |  |  |
| Converged: | 1 | Scale: | 1 |  |  |  |
| No. Iterations: | 14 |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | **Coef.** | **Std.Err.** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | -77.8966 | 4.2903 | -18.1565 | 0 | -86.3054 | -69.4878 |
| **Initial\_days** | 1.4094 | 0.0763 | 18.4837 | 0 | 1.26 | 1.5589 |
| **Lat** | 0.0199 | 0.0173 | 1.1563 | 0.2475 | -0.0139 | 0.0538 |
| **Initial\_admin\_Emergency\_Admission** | 2.3833 | 0.2662 | 8.9516 | 0 | 1.8615 | 2.9051 |
| **Initial\_admin\_Observation\_Admission** | 0.7236 | 0.2688 | 2.6917 | 0.0071 | 0.1967 | 1.2505 |
| **Stroke\_Yes** | 1.5632 | 0.2606 | 5.9984 | 0 | 1.0524 | 2.074 |
| **Arthritis\_Yes** | -1.245 | 0.2162 | -5.7594 | 0 | -1.6687 | -0.8213 |
| **Diabetes\_Yes** | 0.3277 | 0.2192 | 1.4948 | 0.135 | -0.102 | 0.7573 |
| **Anxiety\_Yes** | -0.9108 | 0.2135 | -4.2653 | 0 | -1.3293 | -0.4923 |
| **Asthma\_Yes** | -1.28 | 0.2253 | -5.6813 | 0 | -1.7216 | -0.8384 |
| **Services\_CT\_Scan** | 1.5156 | 0.3401 | 4.4558 | 0 | 0.8489 | 2.1822 |
| **Services\_MRI** | 2.626 | 0.475 | 5.5281 | 0 | 1.6949 | 3.557 |
| **Complication\_risk\_Low** | -1.6417 | 0.2509 | -6.5445 | 0 | -2.1334 | -1.15 |

Our model uses a manageable number of variables, but two p-values have crept over our 0.05 alpha level, so if we want to further reduce this model, we can remove those variables.

**Further Reduced Model:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model:** | Logit | **Pseudo R-squared:** | 0.947 |  |  |  |
| **Dependent Variable:** | ReAdmis Yes | **AIC:** | 723.6874 |  |  |  |
| **Date:** | 9/19/21 15:30 | **BIC:** | 803.0012 |  |  |  |
| **No. Observations:** | 10000 | **Log-Likelihood:** | -350.84 |  |  |  |
| **Df Model:** | 10 | **LL-Null:** | -6572.9 |  |  |  |
| **Df Residuals:** | 9989 | **LLR p-value:** | 0 |  |  |  |
| **Converged:** | 1 | **Scale:** | 1 |  |  |  |
| **No. Iterations:** | 14 |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | **Coef.** | **Std.Err.** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Intercept** | -76.5357 | 4.1251 | -18.5536 | 0 | -84.6207 | -68.4506 |
| **Initial\_days** | 1.4005 | 0.0754 | 18.5696 | 0 | 1.2527 | 1.5483 |
| **Initial\_admin\_Emergency Admission** | 2.3585 | 0.264 | 8.9342 | 0 | 1.8411 | 2.8759 |
| **Initial\_admin\_Observation Admission** | 0.6995 | 0.2678 | 2.6117 | 0.009 | 0.1745 | 1.2244 |
| **Stroke\_Yes** | 1.5214 | 0.2573 | 5.9129 | 0 | 1.0171 | 2.0257 |
| **Arthritis\_Yes** | -1.2106 | 0.2137 | -5.6652 | 0 | -1.6294 | -0.7918 |
| **Anxiety\_Yes** | -0.9343 | 0.2132 | -4.3822 | 0 | -1.3521 | -0.5164 |
| **Asthma\_Yes** | -1.2598 | 0.2239 | -5.6274 | 0 | -1.6985 | -0.821 |
| **Services\_CT\_Scan** | 1.5609 | 0.3381 | 4.6166 | 0 | 0.8982 | 2.2236 |
| **Services\_MRI** | 2.6098 | 0.4778 | 5.4624 | 0 | 1.6734 | 3.5462 |
| **Complication\_risk\_Low** | -1.6137 | 0.2487 | -6.4877 | 0 | -2.1012 | -1.1262 |

My initial research question was related to *all* variables, so I included all variables that contained any patient information that could possibly be a predictor of days of initial hospital stay. I left out variables that were nonsensical codes (UID, customer\_id, etc.), or had too many categories to quantify with dummy variables (job), and/or extra categories related to location (county, zip, etc.). The reduced model kept all variables that were both statistically significant (based on a p-value of 0.05) and not correlated with each other (based on VIF values of 10). Variables with p-values above 0.05 or VIF values of above 10 were removed from the initial model. For this final model, I left the constant/intercept in regardless of its VIF because much of the variables are binary and/or categorical, leading to a small variance, and thus increasing the VIF for the constant/intercept. Furthermore, I’m not worried about multicollinearity with something that is not an actual variable in my dataset.

We now have a very good reduced model, with all p-values less than 0.05, meaning all of our independent variables are statistically significant in predicting ReAdmis. Our pseudo R-squared value of 0.947 implies that that 94.7% of our data can be explained by the model, which is a relatively high percentage and strong fit.

**Confusion Matrix:**

Graphical user interface

Description automatically generated

[[1249 13]

[ 12 726]]

This confusion matrix tells us the accuracy of our model, and below, we can see various measures of the accuracy. By splitting and training our dataset, we can test what happens when our model meets a sample of our data (Bronshtein, 2017). With precision, recall, and accuracy scores of over 0.98, this is a very precise model.

**Precision Score**

True positives / (True positives + False positives) = 726 / (726 + 13) = 0.9824 = ***98.24%***

**Recall Score**

True positives / True positives + False negatives) = 726 / (726 + 12) = 0.9837 = ***98.37%***

**Accuracy Score**

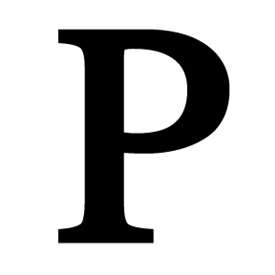
(Correct / All) = (1249 + 726) / (1249 + 726 + 12 + 13) = 0.9875 = ***98.75%***

Supporting code attached in Jupyter notebook.

**Part V: Data Summary and Implications**

*Equation of reduced regression model:*

exp((Initial\_days \* 1.4005) + (Initial\_admin\_Emergency\_Admission \* 2.3585) +

1.2106) - (Anxiety\_Yes \* 0.9343) - (Asthma\_Yes \* 1.2598) + (Services\_CT\_Scan \* 1.5609) + (Services\_MRI \* 2.6098) + (Complication\_risk\_Low \* 1.6137))

= \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1 + exp((Initial\_days \* 1.4005) + (Initial\_admin\_Emergency\_Admission \* 2.3585) + (Initial\_admin\_Observation\_Admission \* 0.6995) + (Stroke\_Yes \* 1.5214) - (Arthritis\_Yes \* 1.2106) - (Anxiety\_Yes \* 0.9343) - (Asthma\_Yes \* 1.2598) + (Services\_CT\_Scan \* 1.5609) + (Services\_MRI \* 2.6098) 1 Complication\_risk\_Low \* 1.6137) - 76.5357

Where P equals the probability of ReAdmis being 1 (meaning the patient is readmitted).

This equation tells us several things. For starters, it means we can indeed predict the probability that a patient will be readmitted, with a high degree of accuracy. The high pseudo R-Squared value of our model (0.947) gives us confidence that our model accurately fits the data. In terms of individual dependent variables, we can make the following statistical assumptions:

* For every additional day of initial hospital visit, readmission is *4.06* times more likely to occur, holding all other variables constant.
* If a patient was an “emergency” admission, readmission is *1.06* times more likely to occur, holding all other variables constant.
* If a patient was an “observation” admission, readmission is *2.01* times more likely to occur, holding all other variables constant.
* If a patient had a stroke, readmission is *4.58* times more likely to occur, holding all other variables constant.
* If a patient has arthritis, readmission is *2.98* times less likely to occur, holding all other variables constant.
* If a patient has anxiety, readmission is *3.92* times less likely to occur, holding all other variables constant.
* If a patient has asthma, readmission is *2.84* times less likely to occur, holding all other variables constant.
* If the primary service the patient received was “CT Scan”, readmission is *4.76* times more likely to occur, holding all other variables constant.
* If the primary service the patient received was “MRI”, readmission is *1.36* times more likely to occur, holding all other variables constant.
* If the patient’s complication risk was “low”, readmission is *1.99* times less likely to occur, holding all other variables constant.

The italicized numbers above were calculated by taking the exponential of the coefficients.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dep. Variable:** | ReAdmis Yes |  |  |  |  |  |
| **Method:** | dydx |  |  |  |  |  |
| **At:** | overall |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | **dy/dx** | **std err** | **z** | **P>|z|** | **[0.025** | **0.975]** |
| **Initial\_days** | 0.0149 | 8.39E-05 | 178.076 | 0 | 0.015 | 0.015 |
| **Initial\_admin\_Emergency Admission** | 0.0252 | 0.002 | 10.199 | 0 | 0.02 | 0.03 |
| **Initial\_admin\_Observation Admission** | 0.0075 | 0.003 | 2.637 | 0.008 | 0.002 | 0.013 |
| **Stroke\_Yes** | 0.0162 | 0.003 | 6.228 | 0 | 0.011 | 0.021 |
| **Arthritis\_Yes** | -0.0129 | 0.002 | -5.936 | 0 | -0.017 | -0.009 |
| **Anxiety\_Yes** | -0.01 | 0.002 | -4.504 | 0 | -0.014 | -0.006 |
| **Asthma\_Yes** | -0.0134 | 0.002 | -5.903 | 0 | -0.018 | -0.009 |
| **Services\_CT\_Scan** | 0.0167 | 0.003 | 4.786 | 0 | 0.01 | 0.023 |
| **Services\_MRI** | 0.0278 | 0.005 | 5.716 | 0 | 0.018 | 0.037 |
| **Complication\_risk\_Low** | -0.0172 | 0.002 | -6.89 | 0 | -0.022 | -0.012 |

The above table tells us the marginal effect of each of our variables (Raoniar 2020). For example, every additional unit of initial\_days will increase the odds of being readmitted by 1.49%, if a patient is an emergency observation, their odds of being readmitted increase by 2.52%, and so on.

The limitations of this data analysis are significant. As I discussed in Task 1, I have considerable skepticism about the accuracy of this data. The bimodality of the initial\_days distribution gives me cause for question. Furthermore, there is no way to validate the data to confirm. It is also important that we do not mistake correlation for causation, and that we understand that just because these variables are correlated with readmission status does not mean they are the cause. Also, the fact that these have been trends in the past doesn’t automatically imply they will be trends in the future.

My recommendation at this point would be to begin using this model to predict whether a patient will be readmitted. Despite the limitations outlined in the paragraph above, this is a strong model that we can use to make predictions. Our ultimate goal is to avoid readmissions and the consequent penalties, so we should input every patient’s information into this model so we can understand their probability of being readmitted. If a patient’s probability is high, we should adjust our care to lower that probability. I’m not a medically trained professional, so I can’t say what that means, but if we pass the information onto the doctors, nurses, and other caregivers, we will be doing our job.

References

Bronshtein, A. (2017, May 17). *Train/Test split and cross validation in Python*. Medium. Retrieved September 22, 2021, from https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6.

Raoniar, R. (2020, October 8). *Fitting MLR and Binary logistic regression using Python*. One Zero Blog. Retrieved September 22, 2021, from https://onezero.blog/fitting-mlr-and-binary-logistic-regression-using-python/.