Lung Cancer: Predictors

Dina Koes and Kellie Halladay

COMP 4447: Data Science Tools 1

Dr. Wojciech Kossek

March 11th, 2024

Research Question

What are some common predictors of lung cancer? Can we build a model to predict the likelihood of being diagnosed with lung cancer based on various health factors, demographics, and habits?

Introduction and Methodology

Lung cancer is one of the leading cause of cancer deaths worldwide and is a top cause of mortality in the US (Howlader, Forjaz, Mooradian, Meza, Kong, Cronin, Mariotta, Lowy, Feuer). While there are screening methods for known factors contributing to lung cancer, such as smoking habits, gender, existing diseases, and environmental factors (Tanoue, Tanner, Gould, Silvestri), there does not appear to be a tool out there to examine all predictors of the disease and determine a person's risk of disease based on these predictors. Such a predictive model would be highly valuable as it would allow people to understand their personalized lung cancer risk based on trends we've seen in demographics, habits, preexisting health factors, and environmental factors. Those identified as high-risk could then take preventative measures, make lifestyle changes to mitigate risks and undergo more frequent screening.

To develop a predictive risk model, two datasets related to lung cancer were obtained from Kaggle for analysis. The first dataset contained comprehensive information on 462,000 participants from China over six years, including demographics, habits, existing health conditions, symptoms, and lung cancer status. The second dataset had records for 309 participants taken from an online lung cancer prediction system with similar information and a lung cancer diagnosis indicator.

To merge the datasets, critical variables like age, gender, smoking, history of chronic disease, fatigue, alcohol use, and symptoms were retained from each dataset. Some adjustments

were necessary to ensure consistency between the datasets, including modifying the gender coding, creating a sequential index variable for the second dataset, and adjusting the variable scales and formats to represent information similarly between the two datasets. After these changes, the datasets could be combined into one consolidated dataset for further analysis of factors related to lung cancer diagnosis. The goal was to integrate the two datasets after making the necessary modifications to have uniformity in variables and formatting between the datasets.

The data preprocessing involved several steps. First, the variables were renamed to have consistent and clear naming conventions. The researchers kept only the columns in both datasets to avoid missing values when combining them. Any unnecessary columns were removed. Next, in the first dataset, an additional variable - LC_Level_numeric - was created, which specifies whether or not the individual has lung cancer, and was set to 1 for every instance in this dataset since this dataset is specific to individuals with lung cancer. Then, in the second dataset, a new gender variable was created, switching M (male) and F (female) to 1 and 2, respectively, which matches the formatting of the first dataset. Next, many of the variables from the first dataset are represented with a number from 1-9 to specify the level. In this dataset, many variables had a value of 1 for ‘No’ or 2 for ‘Yes’. The researchers created new versions of these variables with updated logic that keeps the 1 values unchanged, but changed the 2 values to 6 - a moderate level, according to the other dataset's scaling. A LC_Level_Numeric variable was created, giving a value of 0 for NO LUNG_CANCER and 1 for YES LUNG_CANCER.

The final set of variables were index, age, gender, alcohol use, smoking, shortness of breath, fatigue, chronic lung disease, and lung cancer level numeric. Next, the two dataframes were merged into one consolidated dataframe named lc_data_merge. The Age variable was transformed by taking its logarithm to address scaling differences. Similarly, the Shortness of

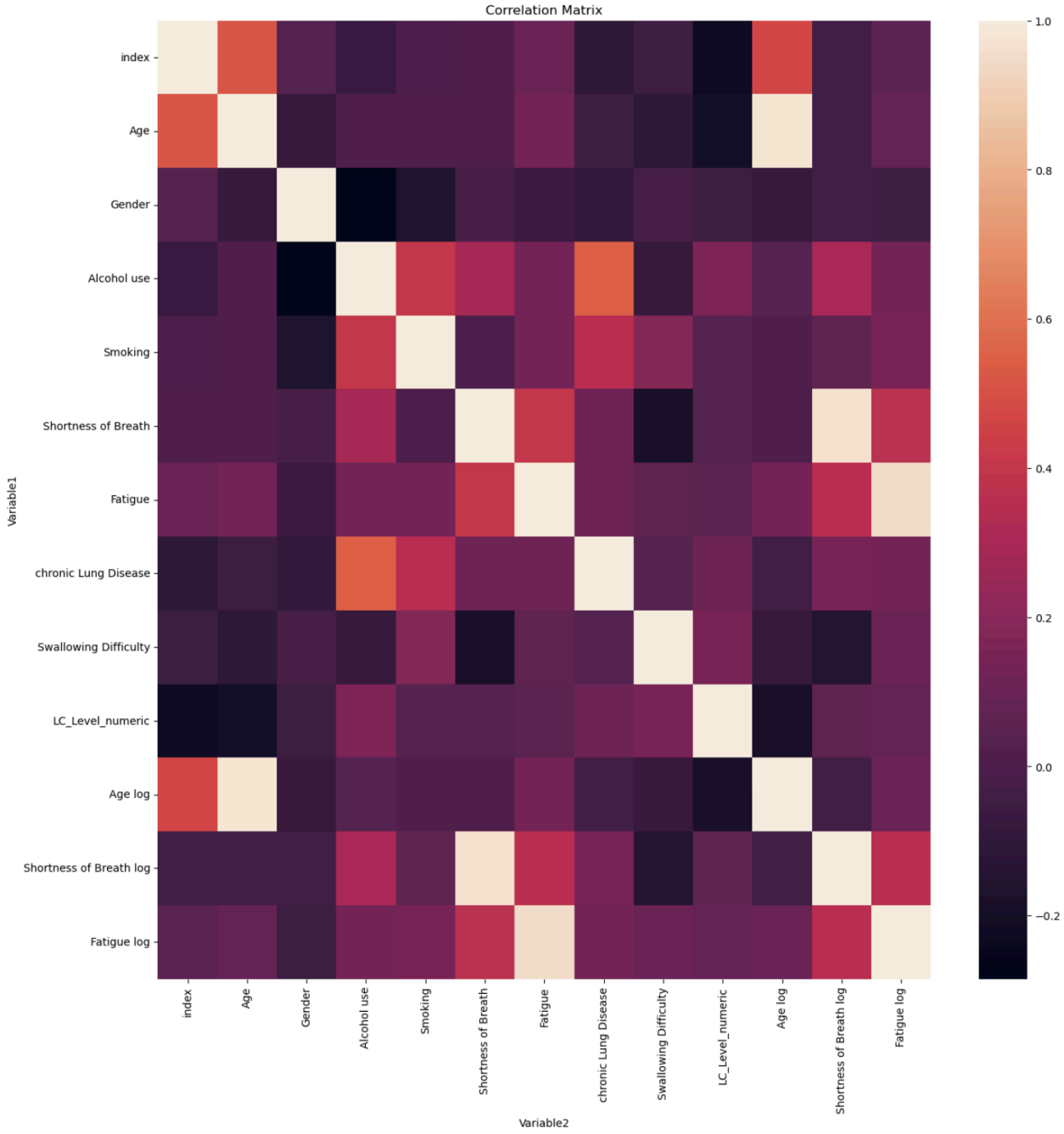
Breath and Fatigue variables underwent logarithmic transformations to correct for skewed distributions. A summary statistics table was generated to showcase descriptive statistics like mean, median, minimum, and maximum for each variable.

Additionally, boxplots were created to inspect the distributions of the variables visually. Comparing the boxplots of the original versus the log-transformed versions of Age, Shortness of Breath, and Fatigue highlights the impact of the logarithmic transformations in correcting skewness and scaling issues. Overall, multiple steps were taken to prepare the data for further analysis, including renaming, removing unnecessary columns, merging, log transformations, and graphical analyses.

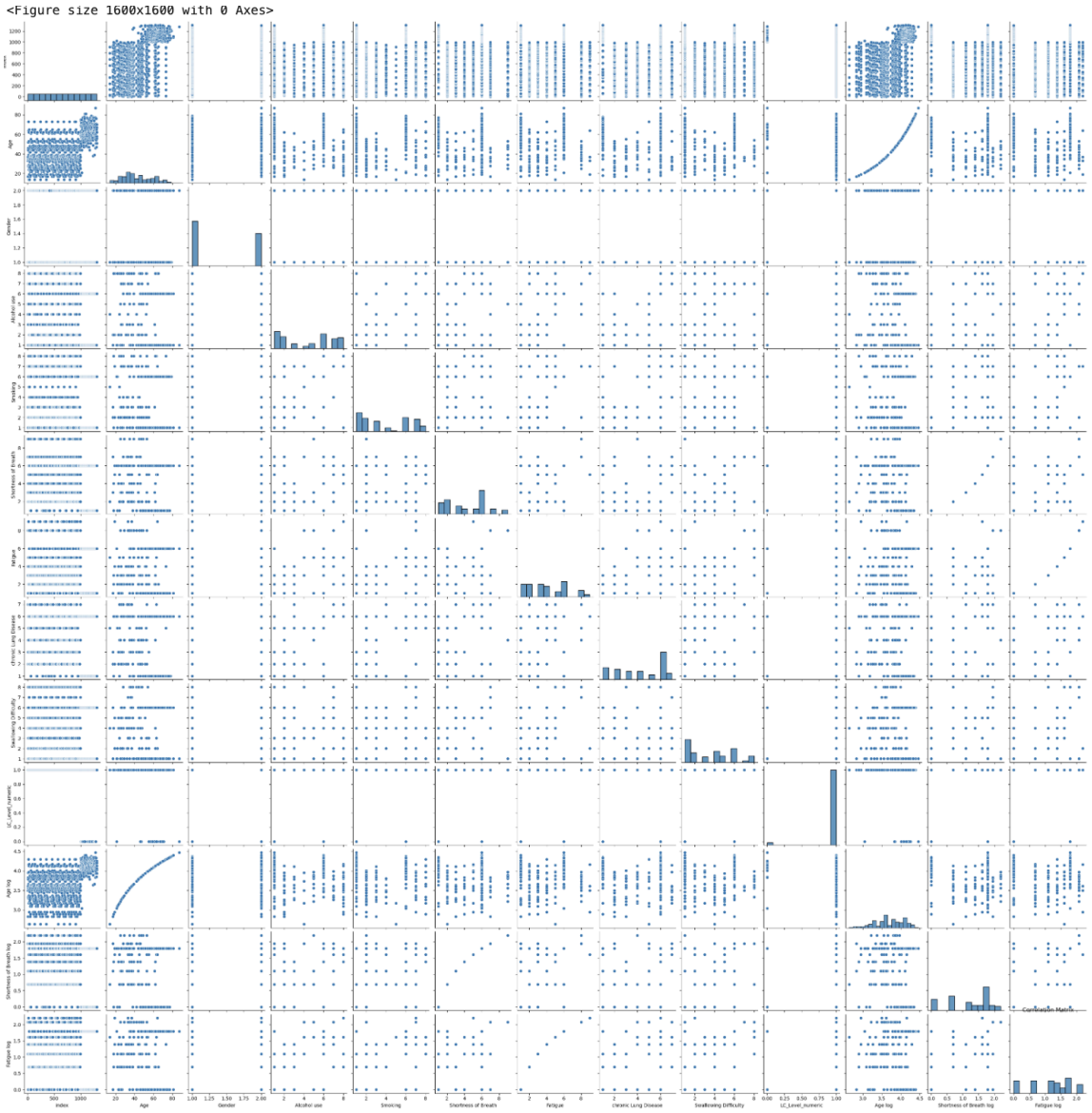
Next, the researchers computed the correlation between the LC_Level_numeric variable, which served as the target variable, and the remaining variables. Upon analysis, the Age variable appeared to have the strongest correlation with the target variable, with a correlation coefficient of approximately -0.19725. Following Age, alcohol use displayed a notable correlations of approximately 0.16668. These findings suggest that age and alcohol use may play significant roles in influencing the likelihood of lung cancer.

	v1	v2	correlation
126	LC_Level_numeric	LC_Level_numeric	1.000000
117	LC_Level_numeric	index	-0.223931
118	LC_Level_numeric	Age	-0.197247
127	LC_Level_numeric	Age log	-0.180174
120	LC_Level_numeric	Alcohol use	0.166675
125	LC_Level_numeric	Swallowing Difficulty	0.151168
124	LC_Level_numeric	chronic Lung Disease	0.118057
129	LC_Level_numeric	Fatigue log	0.080246
128	LC_Level_numeric	Shortness of Breath log	0.059844
119	LC_Level_numeric	Gender	-0.051386
123	LC_Level_numeric	Fatigue	0.041568
121	LC_Level_numeric	Smoking	0.033815
122	LC_Level_numeric	Shortness of Breath	0.031154

A heatmap is a graphical representation of data where colors depict individual values within a matrix. It offers a two-dimensional portrayal of data, making it useful for visualizing distributions, patterns, and correlations. Heatmaps provide an intuitive way to discern trends and variations in the dataset, withdifferent darker colors indicating higher values or correlations, while lighter colors represent lower correlation values.



On the other hand, a pairplot, also known as a scatterplot matrix, is a comprehensive visualization tool used to examine the relationships between pairs of variables within a dataset. Each cell in the matrix represented a scatterplot, illustrating the correlation between the variable represented by the row and the variable defined by the column. Pairplots were instrumental in visualizing the correlation matrix and identifying patterns or trends in the dataset. They allow for a quick and simultaneous comparison of multiple variables, aiding in the exploration and analysis of complex datasets.



To continue, after writing the merged data to a CSV file, the researchers created the final dataset, selecting only the fields necessary for the analysis. These fields included LC_Level_numeric as the target variable and Gender, Age log, Smoking, Chronic Lung Disease, Fatigue log, Alcohol use, Shortness of Breath log, and Swallowing Difficulty as the feature variables.

Once the researchers had the final dataset, they split it into a training and test sets, following a 70-30 split ratio; meaning that 70% of the data was used to train the model, while the remaining 30% was reserved for testing its performance.

The researchers further divided the data into feature sets and target sets. The feature sets contained the independent variables (features) that would be used to predict the target variable (LC_Level_numeric). In contrast, the target sets contained the corresponding actual values of the target variable for both the training and test datasets. This separation was essential for training and accurately evaluating the predictive model's performance.

Linear Regression

The Linear Regression model coefficients represented the weights assigned to each feature variable in the model, indicating the strength and direction of their relationship with the target variable. Positive coefficients suggested a direct relationship, where an increase in the feature variable corresponded to an increase in the target variable, while negative coefficients implied an inverse relationship. Each coefficient's magnitude reflected its influence on the target variable, with larger values indicating more substantial effects. Additionally, the root mean squared error (RMSE) measured the average discrepancy between the predicted and actual values in the dataset. A lower RMSE signified higher accuracy, indicating that the model's predictions were closely aligned with the observed data. Conversely, a higher RMSE suggested more significant

deviation between predicted and actual values, indicating poorer model performance. RMSE was widely utilized as an evaluation metric to assess the goodness of fit of regression models, providing insights into their predictive capabilities.

There are a few additional ways to compute a model's accuracy. AIC is a measure of the goodness-of-fit of a statistical model to data. The aim of the Akaike information criterion (AIC) is to minimize the difference between the observed and predicted values of an outcome variable by selecting the parameters that best explain the variability in the response variable. The goal is to evaluate the balance between explanatory power and model complexity. The formula for calculating AIC is:

- $AIC = -2\ln(L) + 2k$, where L is the likelihood of the data and k is the number of parameters

BIC is a measure of the goodness-of-fit of a statistical model. It stands for "Bayesian Information Criterion". The formula for BIC is:

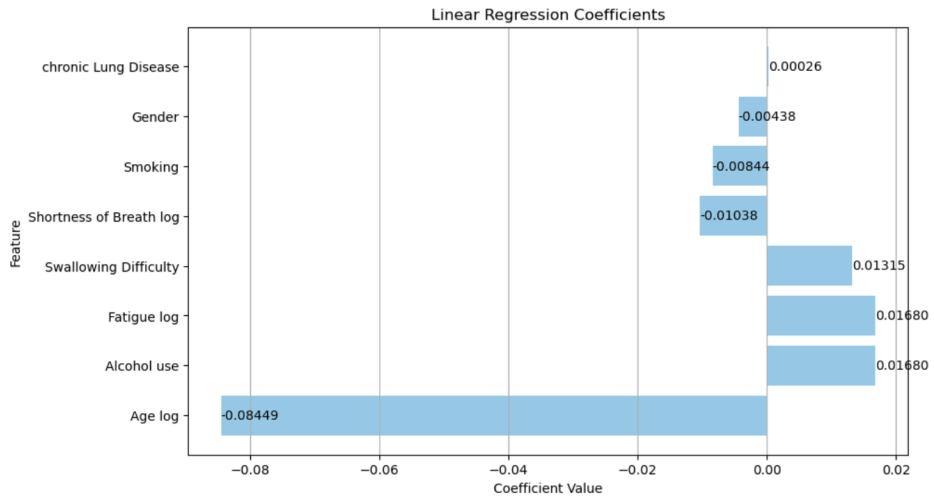
- $BIC = -2\ln(L) + 2k$, where L and k are the likelihood and number of parameters, respectively, and n is the number of records in the test set

Lower AIC and BIC values imply that the model fits the data well.

The Linear Regression model exhibited a relatively low RMSE score of 0.14611, indicating that its predictions closely matched the actual values, and that the model accurately captured the relationship between the dependent and independent variables. The model achieved an AIC score of -378.49605, and the BIC score of -342.73176. These scores suggest that the Linear Regression model is a good fit for the data, as both the AIC and BIC scores are low.

A bar chart was generated to visualize the model's coefficients, displaying each feature alongside its respective coefficient. The chart was sorted based on the absolute magnitude of the

coefficients, facilitating the identification of features with the most significant impact on the model's predictions. This visualization aided in understanding the relative importance of each feature in the Linear Regression model.



Decision Tree Regression

The Decision Tree Regression model operates by recursively partitioning the dataset based on feature values, making decisions at each node to predict the target variable. This approach enables capturing complex, non-linear relationships and feature interactions. With a mean RMSE of 0.09968, the Decision Tree Regressor's predictions deviated from actual values by approximately 0.1 units on average. The RMSE served as a valuable metric for evaluating and comparing the model's performance to other models.

The standard deviation of 0.10222 reflected the variability in RMSE across multiple model runs. A lower standard deviation indicated consistent performance, while a higher standard deviation suggested performance variability potentially influenced by the training data's randomness.

The AIC and BIC values further illuminated the model's fit. The lower AIC (-398.65341) implied a better fit per this criterion, while the higher BIC (-366.86294) raised concerns about

potential overfitting. These insights helped gauge the model's effectiveness and guide adjustments to optimize its performance and generalizability.

Random Forest Regression

A Random Forest Regression model is a powerful ensemble method that combines multiple decision tree classifiers to improve predictive accuracy and mitigate overfitting. Random Forest Regression models achieve robustness and generalizability by fitting Decision Tree Regressions on different subsets of the dataset and aggregating their predictions. This approach reduces variance and enhances performance on unseen data by averaging predictions across multiple trees.

With a mean RMSE of 0.10253, the Random Forest Regressor's predictions deviated from actual values by approximately 0.1 units on average. The RMSE served as a reliable metric for evaluating and comparing model performance with other models.

The low standard deviation of 0.06266 indicated the variability of the RMSE over multiple runs of the model. This value can be used to assess the consistency of the model's performance. A low standard deviation suggests that the model's performance is consistent, while a high standard deviation suggests that the model's performance is variable and may depend on the specific random sample used for training.

The AIC and BIC values for the Random Forest Regression model revealed insights into its fit; it was observed that the AIC value (-534.18201) is higher than the BIC value (-502.39153). This discrepancy in values indicates that the Random Forest Regression model might be a better fit for the data according to the BIC criterion, suggesting that it is penalizing complexity more heavily than the AIC. However, the higher AIC value implies that the model may be overfitting the data according to the AIC criterion, as it favors goodness of fit but

penalizes model complexity less severely compared to BIC. These findings helped assess the model's effectiveness and guide adjustments to optimize its performance and generalization capabilities.

Logistic Regression

The Logistic Regression model employs the logistic function, also known as the sigmoid function, to predict the probability of a binary outcome rather than the actual outcome itself. This function transforms the linear combination of input features into a probability value between 0 and 1. The coefficients in the model represent the effect of each input feature on the log odds of the output variable. A positive coefficient indicates that an increase in that input feature is associated with higher log odds of the output, while a negative coefficient suggests the opposite relationship. For example, in our model, a one-unit increase in the log of Age decreases the log odds of the output by 2.60603, while a one-unit increase in Fatigue log raises the log odds by 0.58963. The RMSE of 0.15132 reflects the average difference between the model's predicted and actual values, implying a reasonable fit to the data. The AIC value is -350.91223 and the BIC value is -315.14794 for the Logistic Regression model. These values suggest that the model adequately explains the variability in the data and provides a parsimonious representation of the underlying relationship. The negative values suggest that the model has a good fit, and the low BIC value further supports this by considering model complexity.

Polynomial Regression - Degree 2

A Polynomial Regression model of Degree 2, or Quadratic Regression, uses a second-degree polynomial equation to model nonlinear relationships between dependent and independent variables. The model incorporates an intercept term, a linear term, and a quadratic term, represented by coefficients b0, b1, and b2 respectively, for each variable. Unlike Linear

Regression models, Polynomial Regression can capture nonlinear associations approximated by a parabolic curve. To make predictions using this model, the input x value can be plugged into the equation along with the coefficient values. For example, using the provided coefficients, when x is 1, the predicted y value is 0.034. The RMSE of 0.14400 for this model indicates that, on average, the predictions deviate 0.14 units from the actual values, providing insight into the model's accuracy. Evaluation metrics including AIC (-315.94692) and BIC (-133.15167) are low, however, these scores are not as low as the previous model's scores, suggesting the quadratic model is not the best fit for this dataset based on both criteria. In summary, the Degree 2 Polynomial Regression leverages a quadratic equation to accommodate nonlinear relationships but does not make the most accurate predictions for this data based on key evaluation metrics.

Polynomial Regression - Degree 3

A Polynomial Regression model of Degree 3, called cubic regression, uses a third-degree polynomial equation to model nonlinear relationships between the dependent and independent variables. The model includes an intercept, linear, quadratic, and cubic term represented by coefficients b0, b1, b2, and b3 for each variable. Unlike Linear Regression, Polynomial Regression can fit nonlinear patterns approximated by a cubic curve. To make predictions, the x input value can be plugged into the equation along with the coefficients. For instance, with x=1 and the given coefficients, the predicted y value is approximately -5.16. The RMSE of 0.14049 indicates the average deviation between predicted and actual values, providing insight into model accuracy. The AIC (-95.33637) is relatively low, but BIC (564.31602) is quite high, which suggests the cubic model does not suitably fit the data. In summary, the Degree 3 Polynomial Regression leverages a cubic equation to model nonlinear relationships but does not make reasonably accurate predictions for this dataset as evaluated by key metrics.

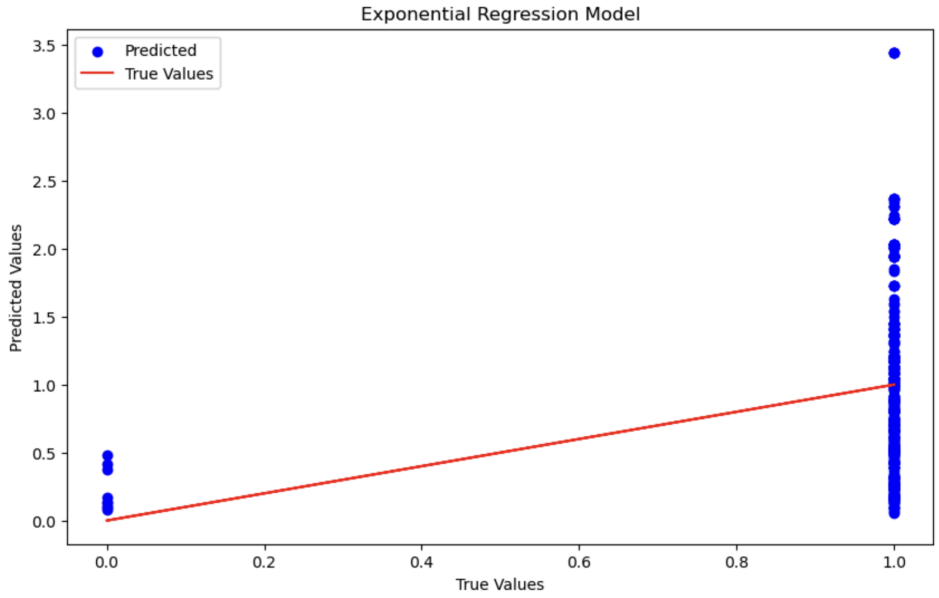
Exponential Regression

The Exponential Regression model is utilized when the relationship between the dependent and independent variables demonstrates an exponential pattern. In this model, the dependent variable y is expressed as a function of the independent variable x through the equation $y=bx$, where b represents the base of the exponential function. This type of model is advantageous when the association between the variables is nonlinear with an exponential trend rather than linear.

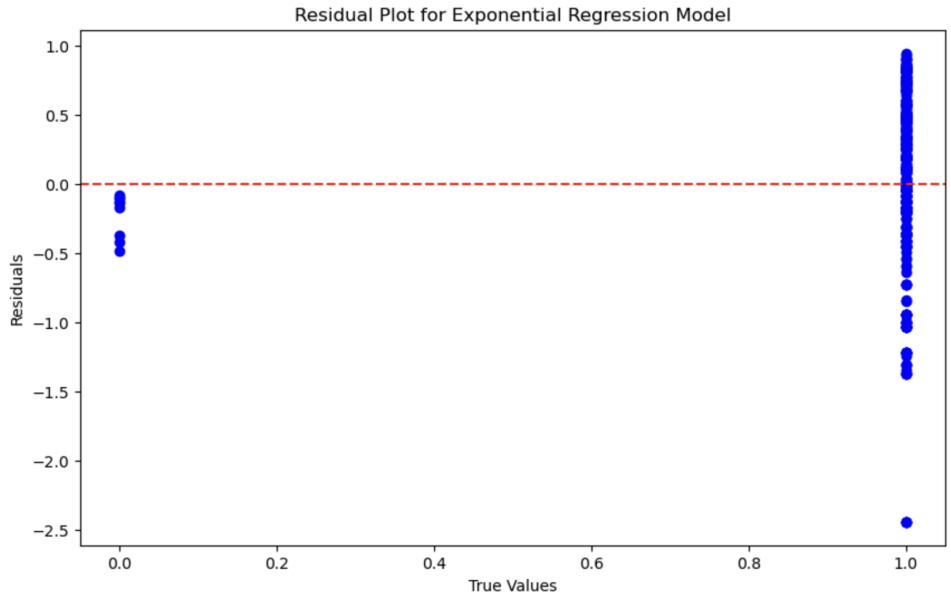
The RMSE of 2.37898 provides the average difference between the predicted and actual values, measuring the model's accuracy. The AIC has an value of 1814.48983 and a BIC value of 1850.25412 for the Exponential Regression model. Given the relatively high AIC and BIC values for the Exponential Regression model, researchers viewed residual values as they can provide valuable insight into the model's performance and assess assumptions, model fit, outliers and prediction accuracy.

A scatter plot comparing true and predicted values visually assesses the model's performance. The residuals or errors signify the differences between the predicted and observed values. A positive residual means the prediction was too high compared to the real value. A negative residual means the prediction was too low compared to the real value. In the data provided, there are two examples that illustrate this. The first example, with index 1148, has a negative residual of -0.25849. This means the model's prediction was slightly higher than the actual observed value. The second example, with index 1049, has a much larger negative residual of -14.297. This indicates the model's prediction was significantly lower than the real observed value for this data point

For example, the first case exhibits a slightly positive residual reflecting overestimation. The second case shows a substantially negative residual, meaning underestimation.



Plotting the residuals facilitates the identification of any systematic patterns or deviations from the ideal zero error line representing perfect prediction. In summary, the Exponential Regression model appears suitable for modeling the exponential relationship in this data based on the accuracy metrics and residual analysis.



Model Summary

The RMSE scores serve as a crucial metrics for assessing the predictive accuracy of regression models. In this analysis utilizing lung cancer predictors, the Decision Tree Regressor exhibited the lowest RMSE score of approximately 0.09968, indicating better accuracy compared to other models like Linear Regression (RSME 0.14611), Logistic Regression (RMSE 0.15133), and Exponential Regression (RMSE 2.37898).

Furthermore, the Random Forest Regression model displayed the lowest AIC score of approximately -534.18201, suggesting a better balance between goodness of fit and model complexity compared to other models. The Linear Regression model closely followed with an AIC score of approximately -378.49605, indicating an effective model. On the other hand, the Exponential Regression model displayed a significantly higher positive AIC score of approximately 1814.48983, indicating a poor fit compared to other models.

Similarly, when considering the BIC criterion, the Random Forest Regression model also emerged as a top performer with the lowest BIC value of approximately -502.39153. This indicates a favorable balance between model fit and complexity. The Linear Regression model following closely with a BIC value of around -342.73176, suggesting a simpler, yet, effective model compared to the others.

Overall, both AIC and BIC criteria support the Random Forest Regression model as the top-performing model, closely followed by the Linear Regression model, for this particular dataset and analysis. Its combination of high predictive accuracy, balanced complexity, and optimal fit to the data solidifies its position as the top-performing regression model among those evaluated.

Using the Model

Now that a preferred model has been chosen based on the various accuracy and goodness of fit scores, it can be used by individuals to test their risk of developing lung cancer based on their specific demographics, habits, health, health, etc. For example, if a 50-year-old male who occasionally smokes and experiences shortness of breath when he does, drinks alcohol a few times a week, regularly feels fatigued, and has no history of chronic lung disease or difficulties with swallowing were uses this model, the inputted values would be as follows:

- gender = 1 for male
- age = 50
- smoking = 3 for the occasional smoke
- cld = 1 for no history of chronic lung disease
- fatigue = 4 for moderate fatigue
- alcohol = 5 for regularly consuming alcohol
- shortnessBreath = 2 for occasionally
- swallowingDiff = 1 for none

The process of determining the likelihood of an individual developing lung cancer involves analyzing various factors such as age, fatigue, and shortnessBreath variables. The log of these variables are computed to generate a comprehensive risk assessment. In this case, after inputting the relevant information into the predictive model, a cancer risk score of 0.92 was obtained on a scale of 0 to 1, indicating that this user is likely to develop lung cancer in his lifetime. This risk score serves as a valuable tool in assessing the potential health outcomes for an individual and can aide in early detection and preventative measures to mitigate the risk of developing lung cancer.

References

Tanoue, L., Tanner, N. T., Gould, M. K., & Silvestri, G. A. (2015). Lung cancer screening. *American Journal of Respiratory and Critical Care Medicine*, *191*(1), 19–33. https://doi.org/10.1164/rccm.201410-1777ci

Howlader, N., Forjaz, G., Mooradian, M. J., Meza, R., Kong, C. Y., Cronin, K. A., Mariotto, A. B., Lowy, D. R., & Feuer, E. J. (2020). The effect of advances in Lung-Cancer treatment on population mortality. *The New England Journal of Medicine*, *383*(7), 640–649. https://doi.org/10.1056/nejmoa1916623