Survival Model for Non-Small Cell Lung Cancer

...

By Joseph Aquino

The Problem

- Lung cancer is the leading cause of cancer-related deaths worldwide, with an estimated 154,050 deaths in the US alone in 2018.
- Prognostic accuracy of life expectancy after diagnosis with NSCLC (non-small cell lung cancer) requires considerable improvement, as genomic data is still only used sparingly.
- Without accurate life expectancy forecasts, healthcare professionals, patients, and their families have significant difficulty determining the best course of treatment and conducting end-of-life decision making.

BUT ... Lung cancer survival time can be predicted with survival regression

There are clinical and genomic features that influence the survival time of NSCLC patients which can be identified by using a survival regression model.

We need a survival regression model that...

- Can predict one-year survival using both clinical and genomic data.
- Can simultaneously relate multiple risk factors to survival time.
- Is capable of handling right-censored survival times (when a patient does not die during the study).
- And will only include features that have a significant effect on survival time.

The most elegant and well-known model for our use is

Elastic Net Penalized **Cox** Proportional Hazard Regression

Using my model

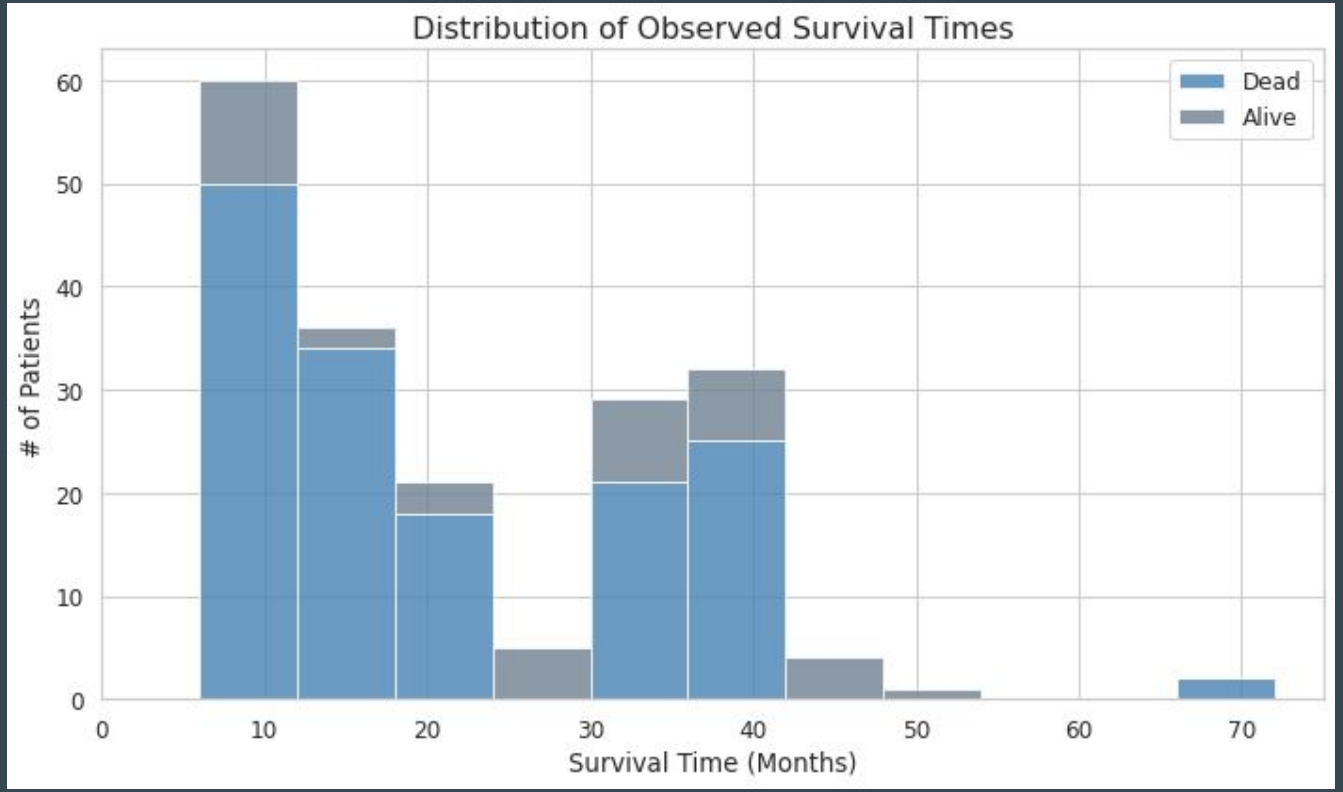
We can predict with relatively good accuracy...

- The outcome (dead or alive) at time t (including t = 1 year) of each test set patient using only 17 clinical and genomic features.
- The survival time after diagnosis of each test set patient.
- The outcome at time t and survival time of any patient diagnosed with non-small cell lung cancer who has data for each of the model features, since the performance of my model has shown to be generalizable.

Observations from Exploratory Data Analysis

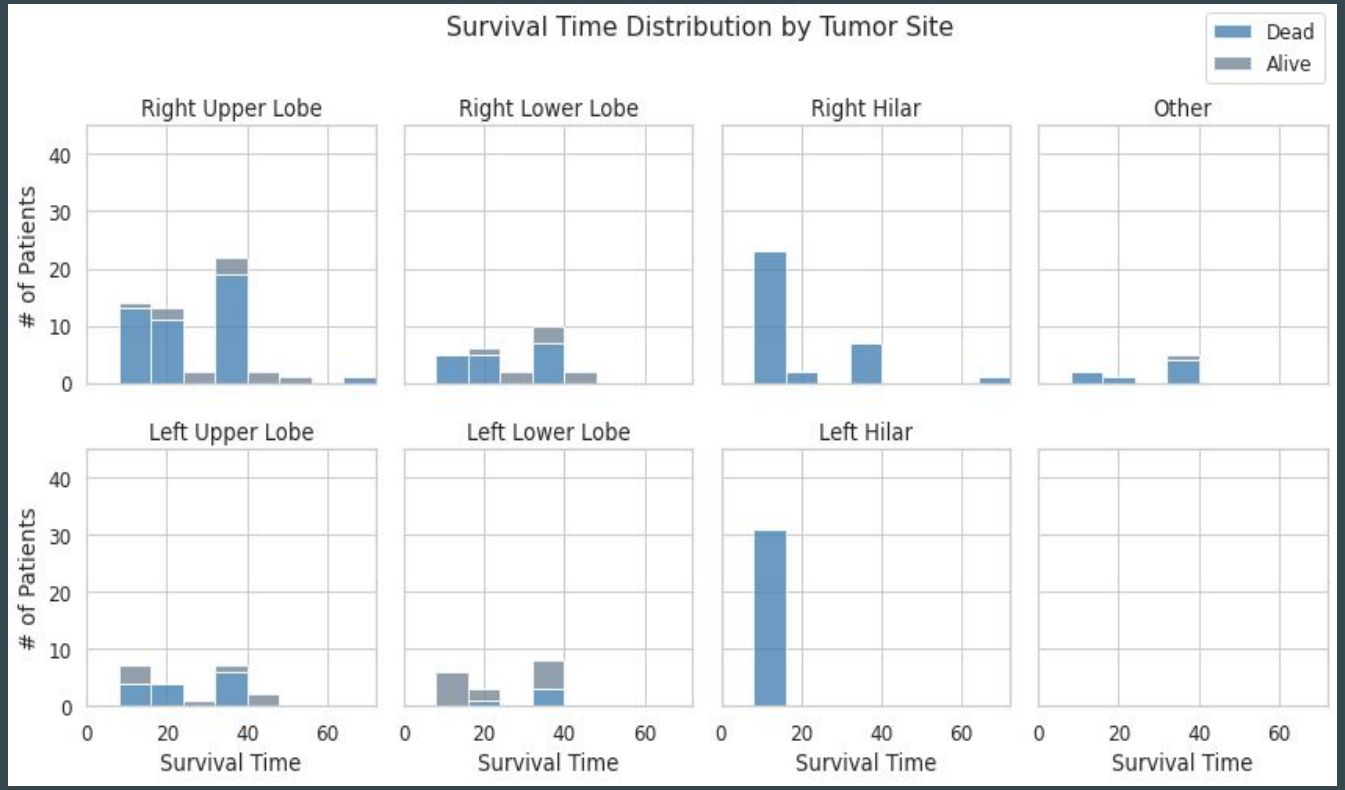
Observed survival times

- Most of the patients seem to fall into 2 groups for observed survival time: approximately 1 and 3 years.
- Analysis of the survival times is complicated by right-censoring (patients alive at follow-up).



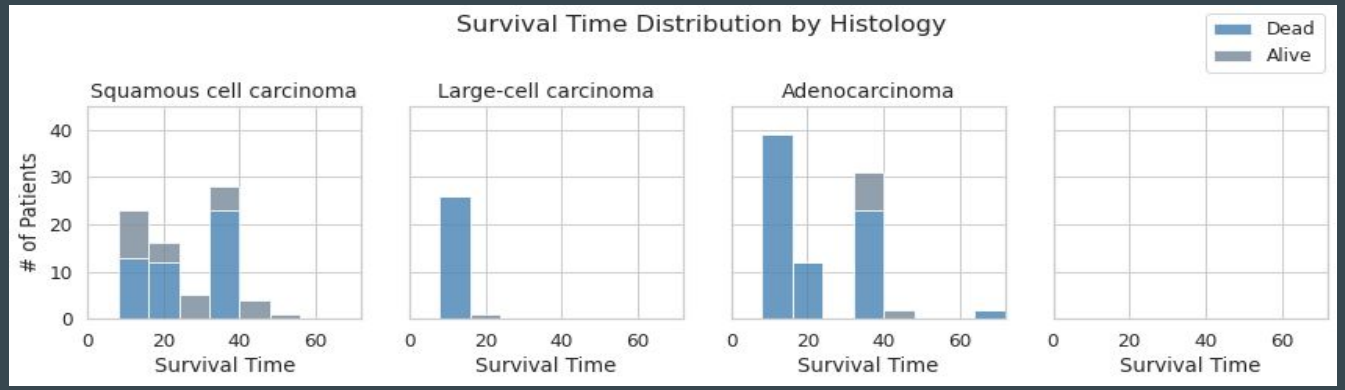
Survival times for different patient groups

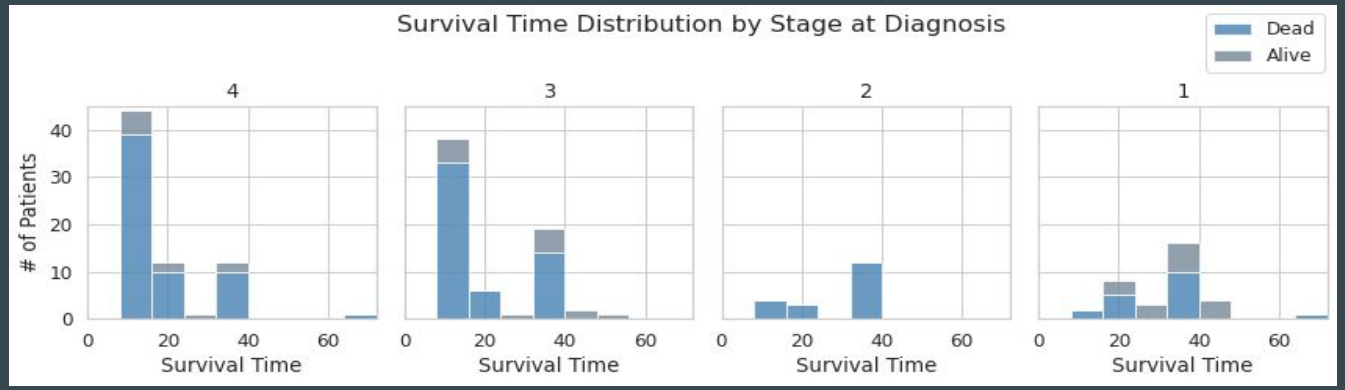
- The following charts show how the survival times of patients differ for each value of the 5 categorical features: *Tumor Site, Histology, Stage at Diagnosis, Tumor Grade, and Tumor Stage.*



Survival times for different patient groups

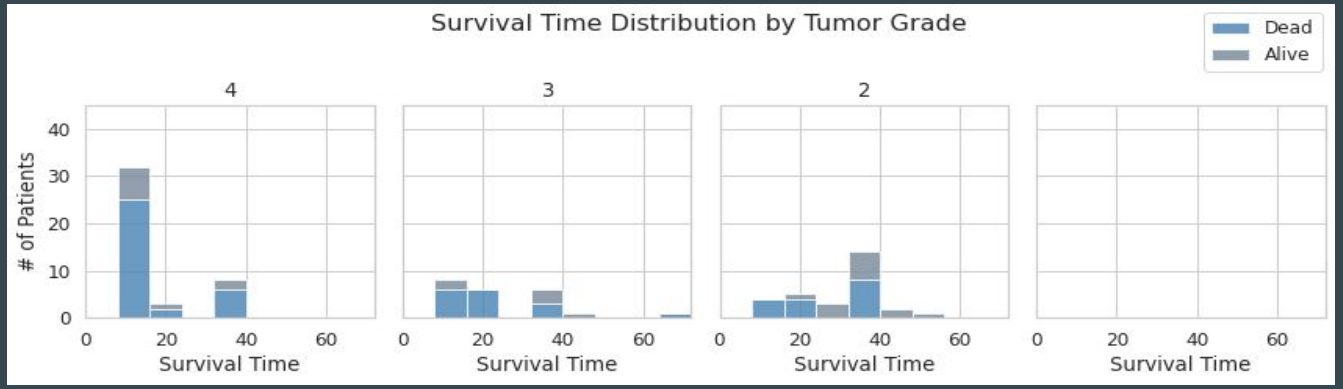
- Large-cell carcinoma histology and hilar tumor site are associated with reduced survival time.
- Conversely, patients with squamous cell carcinoma histology tend to survive longer.

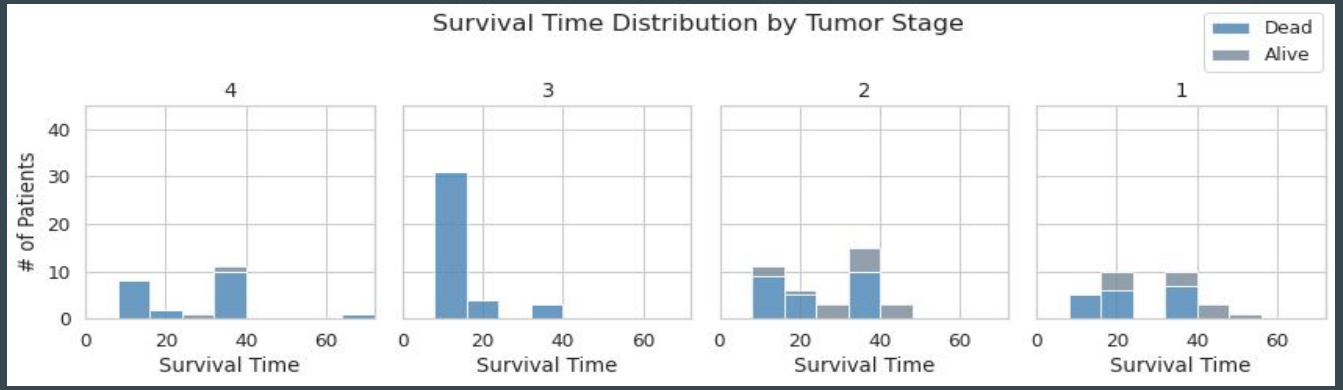




Survival times for different patient groups

- As expected, there is also a clear trend where survival time decreases as cancer stage and grade increase.





Features correlated with survival time

- Clinical features negatively correlated with survival time: *Tumor Stage, Stage at Diagnosis, Left Hilar Tumor Site, Large-cell Carcinoma Histology, and Number of Metastases to Lymph Nodes.*
- Many of these clinical features are also positively correlated with each other.
- Genomic features negatively correlated with survival time: *Mutations of TSC2, MSH2, and APC.*
- These genomic features are also positively correlated with many of the clinical features listed above.
- No features have a significant positive correlation with survival time.

Summary of observations

- Most of the NSCLC patients seem to live for either approximately 1 or 3 years.
- The following features are associated with reduced survival time: *Large-cell Carcinoma Histology, Hilar Tumor Site, and Mutations of TSC2, MSH2, and APC.*
- *Squamous Cell Carcinoma Histology* is the only feature that seems to be associated with increased survival time.
- Survival time tends to decrease as the following features increase: *Tumor Stage, Stage at Diagnosis, and Number of Metastases to Lymph Nodes.*
- Lastly, many of the features negatively correlated with survival time are positively correlated with each other.

Survival Analysis

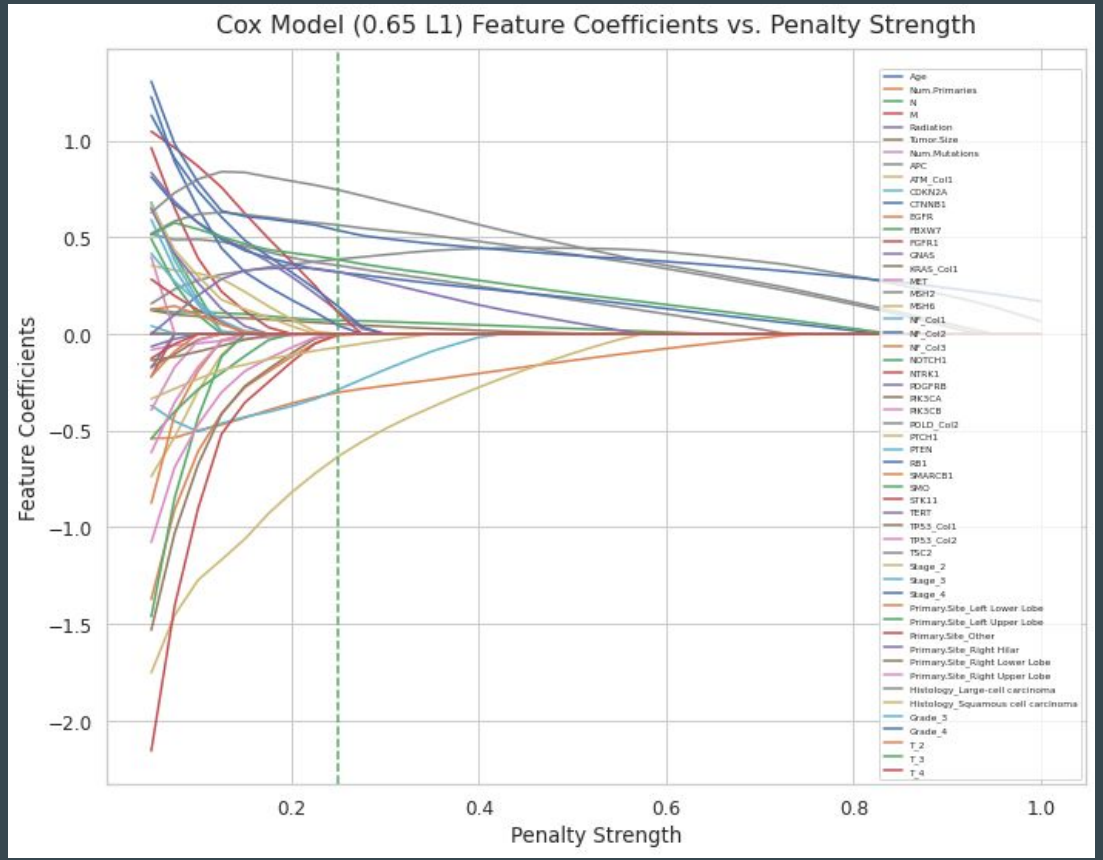
Feature Selection

Feature selection allows you to pinpoint the most relevant features, reduce model complexity, and increase the generalizability of your model.

- My dataset initially contained 63 clinical and genomic features.
- Genomic features (gene mutations) that occurred in less than 2% of the patients were removed, which left 43 features remaining.
- Finally, the elastic net penalty of the survival regression model (Cox proportional hazard regression) reduced many of the feature coefficients to zero, performing embedded feature selection and leaving my final model with just 17 features.

Feature Selection

- During model tuning, the 2 hyperparameters of the elastic net penalty (L1 ratio and penalty strength) were adjusted to create the optimal model.
- The chart shows how elastic net penalization reduces the feature coefficients of the Cox model as penalty strength is increased.

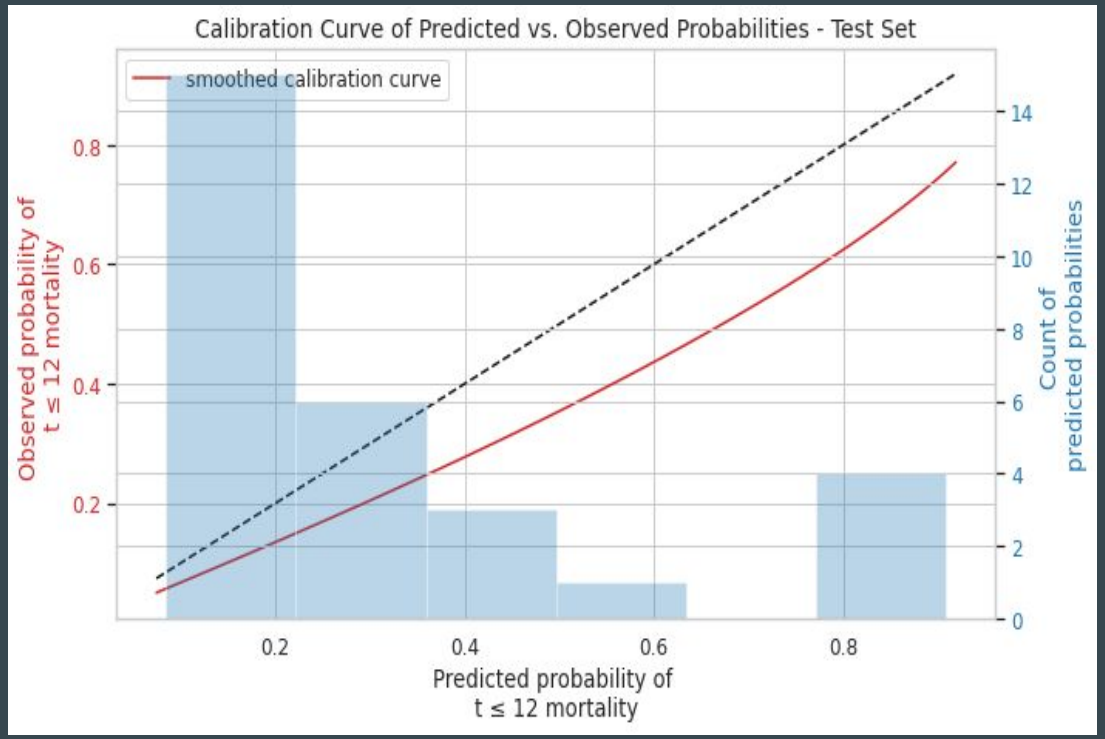


Model results

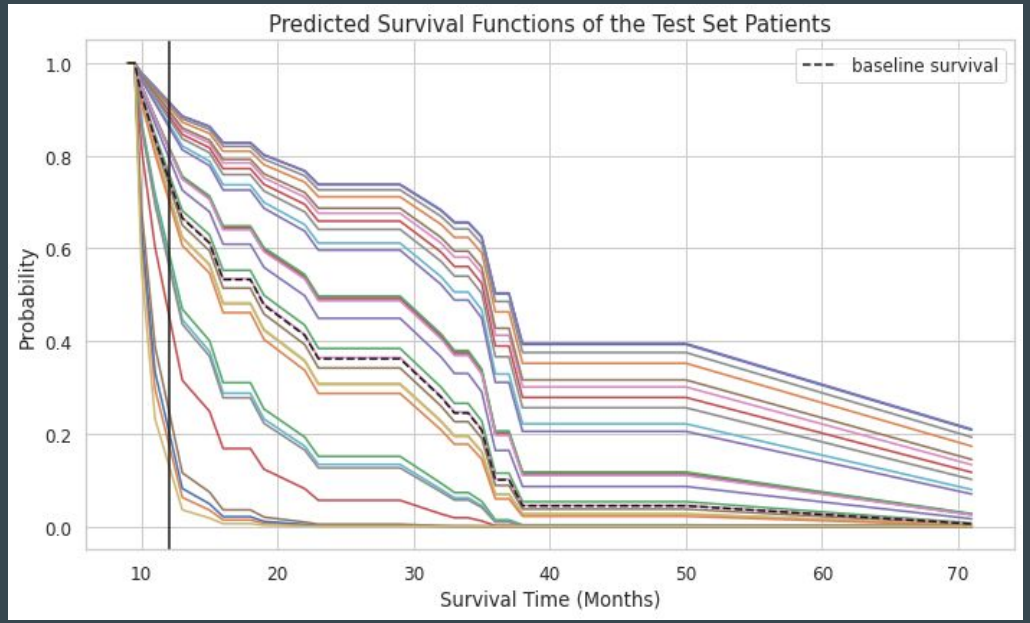
- Our goal is to predict the survival times of patients diagnosed with non-small cell lung cancer with an emphasis on one-year survival.
- To assess the performance of my tuned model on unseen data, the test set, I used the concordance index.
- The concordance index evaluates the accuracy of the ranking/order of predicted survival times of patients. Fitted survival models typically have a score between 0.55 and 0.75.
- My model has a test set concordance index of 0.77. This is an excellent score for a survival model, and indicates that the model likely has generalizable predictive power.

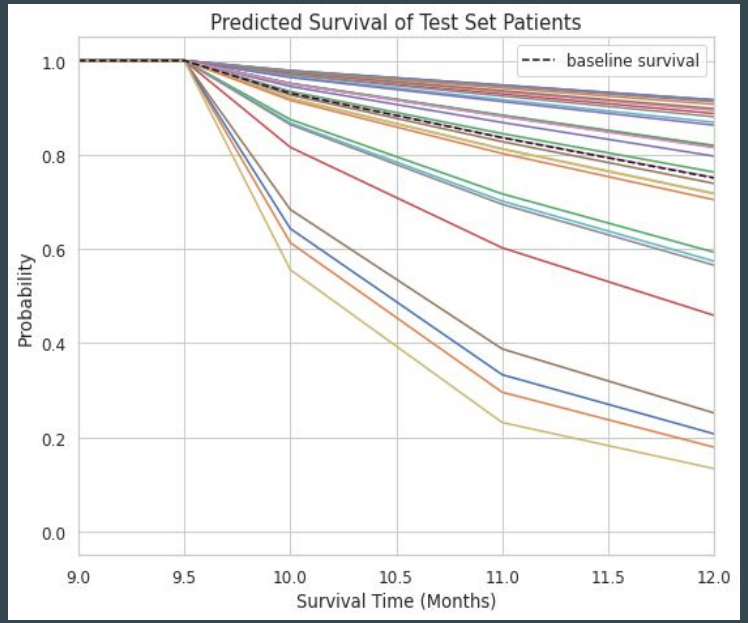
Model results

- This survival probability calibration plot compares the predicted one-year mortality probabilities of my model against the observed probabilities.
- The predicted probabilities of my model are only slightly higher than those observed, indicating that the model is well-calibrated for the one-year timeframe.



Model results

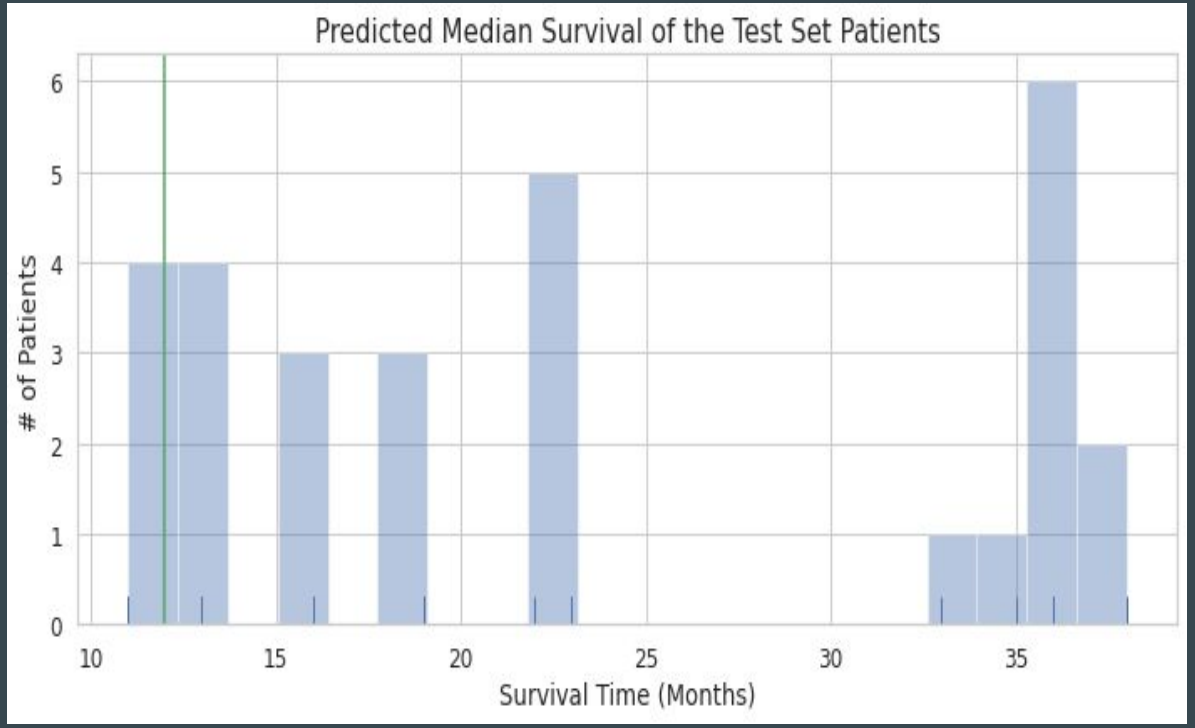




- A survival function gives the probability that a patient will survive beyond time t.
- The one-year baseline survival probability of my model is approximately 75%, while the median baseline survival time is approximately 18 months.
- The one-year survival probabilities of the test set patients range from approximately 15% to 90%.

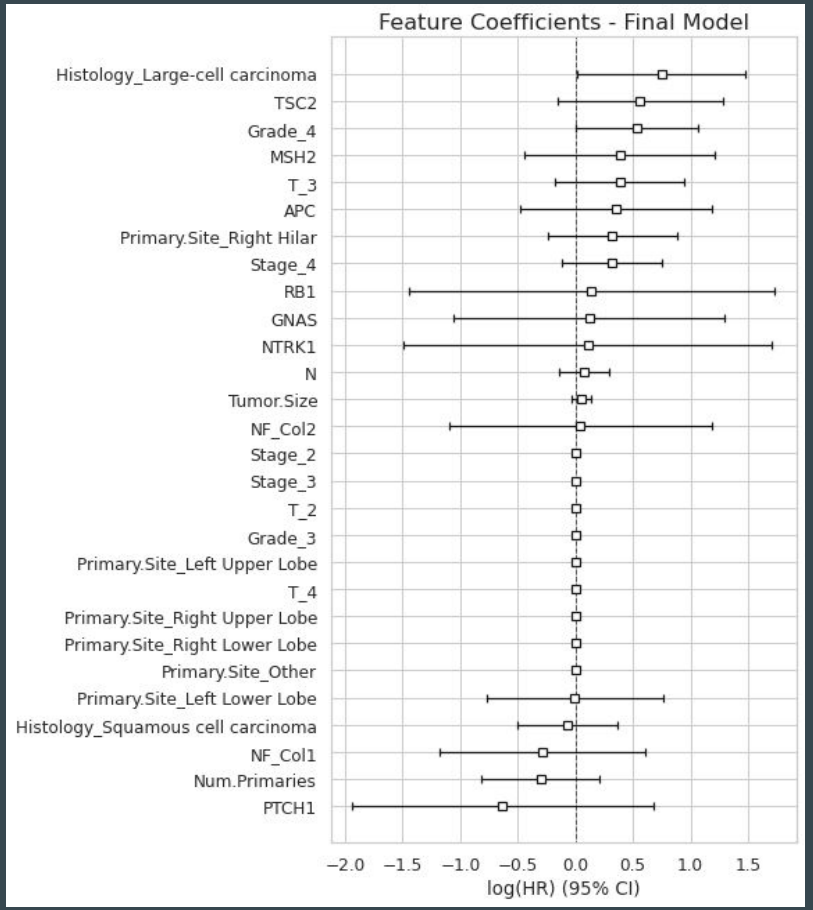
Model results

- The model not only makes probabilistic predictions with survival functions, but it predicts the precise survival time after diagnosis.
- The median survival predictions of the test set patients look very similar to the observed survival times of the dataset as a whole.



Feature importance

- The plot shows the features and feature coefficients of my Cox proportional hazard model.
- A rough estimate of feature importance is given by the absolute value of the feature coefficient.
- The final model has 17 features, although the 5 categorical features are represented by 11 indicator variables (i.e. Grade_4 represents the value Grade 4 of the feature Tumor Grade).



Feature importance

The following is an explanation of hazard ratios associated with Cox regression models:

- The hazard ratio is probably the best metric by which to assess the importance of a feature (unit-scaled if numeric) in relation to others.
- The hazard function/rate gives the probability of death occurring at time t, given it has not occurred yet.
- For a given binary or unit-scaled feature x, the hazard ratio is the ratio of the hazard rate of a patient with x = 1 to the hazard rate of a patient with x = 0 at all times, assuming all other features are the same.
- Meanwhile, for a categorical indicator variable x, the hazard ratio is the ratio of the hazard rate of a patient in the category that x represents, to the hazard rate of a patient in the reference category (the dropped category), assuming all other features are the same.
- For a feature with coefficient b, the hazard ratio of the feature is given by e^b.
- Suppose the hazard ratio of a genomic feature (genetic mutation) is 1.5. Then a patient with the mutation has a hazard rate 1.5 times that of a patient without the mutation at all times, assuming all other features are the same.

Feature importance

- The hazard ratios of the categorical and non-binary numeric features need to be adjusted so that they can be compared more appropriately to those of the binary features.
- The non-binary numeric features are not scaled, and therefore their coefficients and resulting hazard ratios vastly underestimate the influence they have on the model.
- I used the equation of the hazard function of the Cox regression model to calculate the coefficients that correspond to the unit-scaled values of the 2 features.
- The coefficient of Tumor Size increased from 0.052 to 0.465, while that of N increased from 0.070 to 0.209.
- For each categorical feature I decided to calculate the largest hazard ratio among the values of the given feature, and assign that hazard ratio to the feature as a whole.

Feature ranking

By adjusted hazard ratio

1. Histology: 2.25
2. PTCH1: 0.53 → 1.88
3. TSC2: 1.76
4. Tumor Grade: 1.71
5. Tumor Size: 1.59
6. MSH2: 1.47
7. Tumor Stage (T): 1.47
8. APC: 1.42
9. Tumor Site: 1.39
10. Stage at Diagnosis: 1.38

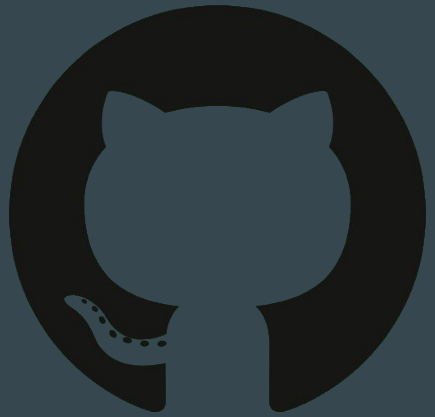
- Almost all of the most influential features were first identified during EDA.
- Each categorical feature has one value with an outlying hazard rate much higher than those of all the other values.
- They are as follows: Large-cell Carcinoma Histology, Grade 4 Tumor Grade, T 3 Tumor Stage, Right Hilar Tumor Site, and Stage 4 Stage at Diagnosis.
- Of the four genomic features on the list, PTCH1 is the only gene where the hazard rate is lower when it is mutated. 1.88 is the inverse of its 0.53 hazard ratio.
- The values of Tumor Size went from a [1, 10] range to a [0, 1] range when I scaled them.
- The updated hazard ratio of Tumor Size is interpreted to mean that a patient with a size 10 tumor has 1.59 times the hazard rate of a patient with a size 1 tumor at all times, assuming all other features are the same.
- The hazard ratio listed for Histology is the hazard ratio of Large-cell Carcinoma with respect to Squamous Cell Carcinoma: $e^{.744-(-.069)} = 2.25$

Potential uses for model

Potential uses for model

Personal Profile links

www.linkedin.com/in/joseph-aquino-a2644612



github.com/joe-aquino