## Part 4: Short-Answer Questions

Answer the following based on your work in **Part 3**:

1. Please provide the link to your public **GitHub repository**.

https://github.com/Sheona-Hans/BINF-5507-Materials/tree/main/Assignment/Assignment4

1. Based on your work in Part 3, please fill out the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Strengths** | **Limitations** | **Key Use Cases** | **Your Observations** |
| **Kaplan-Meier** | Simple and intuitive.  Easy to visualize survival distributions.  Suitable for comparing survival between two or more groups.  Can handle censored data effectively. | Does not account for covariates (age, treatment, etc.).  Assumes independence of survival curves.  Limited to univariate analysis (single factor comparison). | Basic survival analysis.  Comparing survival between groups (e.g., treatment, stage).  Visualizing survival curves. | Provides a good overview of survival over time.  The log-rank test shows strong significance between Early vs Advanced stages. |
| **Cox Proportional Hazards** | Can incorporate multiple covariates (age, sex, treatment, etc.).  Flexible in assessing the relationship between continuous and categorical variables.  Efficient for large datasets. | Proportional hazards assumption may not hold (e.g., with age and stage).  Linear assumptions for continuous variables might oversimplify complex relationships. | Modeling survival with multiple covariates.  Estimating hazard ratios for variables.  Testing the impact of covariates on survival (e.g., treatment effects). | Age and Stage failed the proportional hazards assumption, suggesting the model might need modifications.  Provides more nuanced survival predictions by adjusting for covariates. |
| **Random Survival Forests** | No assumptions on proportional hazards.  Can handle non-linear relationships and interactions between covariates.  Excellent for high-dimensional datasets.  Provides feature importance analysis | Computationally intensive and resource-heavy.  Less interpretable compared to Cox models.  Risk of overfitting without proper tuning.  Requires more data for reliable performance. | Complex datasets with many features.  Feature selection and importance analysis.  Predicting survival with non-linear relationships.  Handling non-proportional hazards. | RSF performed well but had a lower C-index than Cox.  The feature importance output is useful for identifying the most predictive variables. |

1. What are the primary differences between Kaplan-Meier (KM) analysis, Cox regression, and Random Survival Forests?

| **Aspect** | **Kaplan-Meier** | **Cox Proportional Hazards** | **Random Survival Forests** |
| --- | --- | --- | --- |
| **Method** | Non-parametric, univariate | Semi-parametric, multivariate | Non-parametric, machine learning (ensemble of decision trees) |
| **Handling Covariates** | Does not handle covariates (only compares groups based on one factor) | Handles multiple covariates (continuous and categorical) | Handles multiple covariates, including non-linear relationships and interactions |
| **Model Assumptions** | Assumes independent survival curves | Assumes proportional hazards (constant hazard ratio over time) | No assumption of proportional hazards |
| **Use Cases** | Group comparison (e.g., survival between treatments, stages) | Modeling survival with multiple covariates and estimating hazard ratios | Predicting survival with complex datasets, handling non-linear relationships, and identifying important features |
| **Interpretability** | Highly interpretable (survival curve, log-rank test) | Moderately interpretable (hazard ratios for covariates) | Low interpretability (decision trees, but feature importance can be extracted) |
| **Complexity** | Simple and easy to apply, but limited to univariate analysis | Moderate complexity (assumes proportional hazards, can handle multiple covariates) | High complexity, powerful prediction, but less interpretable and computationally intensive |

1. What assumptions are made by Cox Proportional Hazards regression? How can these be evaluated?

| **Assumption** | **Evaluation Method** |
| --- | --- |
| Proportional Hazards Assumption | Log-minus-log plot / Schoenfeld residuals test / Time-varying covariates |
| Linearity of Covariates | Martingale residuals plot / Add splines or polynomial terms for continuous covariates |
| Independence Survival Times | Clustered survival models (e.g., frailty models) for correlated observations (same hospital, etc.) |
| NoOmitted Confounders | Include all known relevant covariates / Conduct sensitivity analysis for robustness |
| Non-informative Censoring | Compare covariates between censored and non-censored individuals / Use informative censoring models if necessary |

1. Which method provided the best balance between interpretability and predictive performance?
   * + 1. **Best Balance**: **Cox Proportional Hazards** provides the best balance between interpretability and predictive performance for typical clinical datasets. It allows for understanding how different covariates influence survival times, but also provides **reasonable predictive performance** without requiring high computational power or complex tuning.
       2. **Kaplan-Meier** is highly interpretable but lacks predictive power as it doesn't model individual survival outcomes. It’s great for initial exploratory analysis but isn’t suited for making predictions.
       3. **Random Survival Forests** offers the best **predictive performance** but lacks interpretability. It's a powerful tool for survival prediction when interpretability is less critical (e.g., for research purposes or high-dimensional datasets), but not ideal for clinical or regulatory settings where explanations of predictions are important.

Thus, if **interpretability is prioritized** and can accept moderate **predictive performance**, then **Cox Proportional Hazards** model is likely the best choice. If **predictive accuracy** is the main goal and interpretability is secondary, then **Random Survival Forests** may be the best option.

1. Identify any features that consistently demonstrate predictive power across different methods and highlight their potential clinical significance.

The following features consistently demonstrate predictive power across Kaplan-Meier, Cox Proportional Hazards, and Random Survival Forests:

1. **Age** - A consistent predictor across methods, highlighting the importance of older age as a risk factor for worse survival outcomes.
2. **Stage of Cancer** - The stage is always a key predictor of survival, with early-stage cancer consistently associated with better survival outcomes.
3. **Treatment Modality (Tx Group)** - The type of treatment (ChemoRT vs. RT alone) plays a significant role in survival outcomes, with combination therapy (ChemoRT) generally associated with better survival.