Decision Support Systems

Shivaram Rammohan
202202968@post.au.dk
Department of Computer Engineering
Aarhus University
Aarhus, Denmark

Instructors:

Christian Fische Pedersen & Christian Marius Lillelund

Group 11

Project 2 - Survival Analysis

Contents

Introduction:

Survival analysis is a statistical technique used to analyze the time until an event of interest occurs. In the context of a telecom network, survival analysis can be applied to understand customer churn behavior. Customer churn refers to the phenomenon where customers switch or terminate their subscription or services with the telecom provider. By utilizing survival analysis, we can gain insights into the factors that influence customer churn and predict the probability of customers churning over time.

In this project, we will use a dataset from a telecom network that contains information about customers and their churn status. The dataset includes features such as gender, senior citizen status, partner status, internet service, contract type, and payment method, among others. Our goal is to explore the relationship between these features and customer churn and develop models to predict churn probabilities.

Project description:

The project is done for the survival analysis of customer churn data. The analysis is done for Telco Customers where various parameters are used to aid training the model and hence, we will also see how these models stack among each other in their predictions.

Report structure:

The report will contain the first theory on how the data is obtained and what kind of models have been used and how their accuracy stands against the existing system and the other models that we have implemented against the data collected. The results and observations are then discussed to find the best model for the Survival analysis of the customer and are better at predicting survival of the customer.

Repository:

Material related to this project is stored in a GitHub account where the report and source code can be found. And, in the ZIP file that has been submitted for this project.

Project Background:

The project discusses predicting the survival or the termination of the user using the telecom service. This project deals with multiple factors where they are factored together in multiple models to predict the probability of survival. The prediction usually uses most of the data columns provided in the dataset and then modelled accordingly.

Approach:

This project achieves the problem statement by using the dataset from Telco Churn data which is then preprocessed to train the model such that the model can determine and predict if the user is about to terminate his service or not. The model is then validated with its accuracy or Root mean square error to determine the better model for prediction. The Flow diagram in Fig 1 will be a diagrammatic representation of the approach.

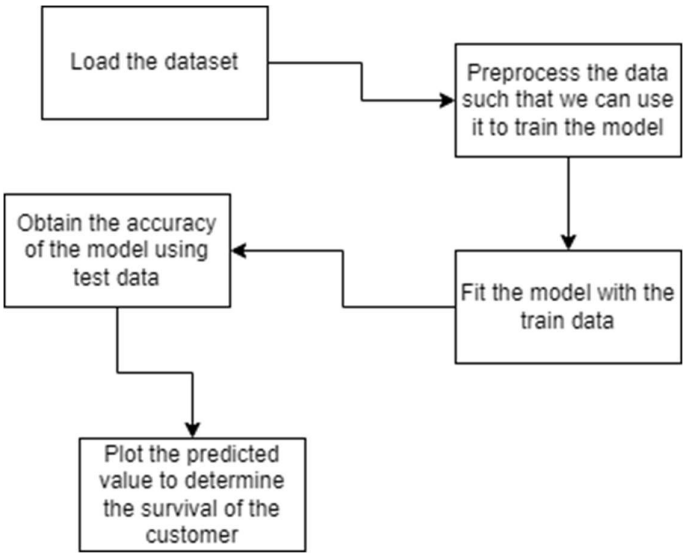

Figure 1: Flowchart for the approach

Dataset and preprocessing:

The dataset that we are using is from Telco which has close to 7000 customers along with a large set of data for each customer like 'customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure' etc. These values are then preprocessed in a onehot encoder to convert into numerical representation of the data in the customer fields as shown in the below code snippet.

```
preprocessing = make_column_transformer(
    (OneHotEncoder(), ['gender', 'SeniorCitizen', 'Partner',
'Dependents', 'PhoneService', 'MultipleLines',
                       'InternetService', 'OnlineSecurity',
'OnlineBackup', 'DeviceProtection', 'TechSupport',
                       'StreamingTV', 'StreamingMovies', 'Contract',
'PaperlessBilling', 'PaymentMethod']),
    remainder='passthrough')
```

Training models and their Results:

Kaplan Meier Method:

The Kaplan-Meier method, also known as the product-limit estimator, is a non-parametric statistical technique used to estimate the survival function of a population over time. It is commonly employed in survival analysis to analyze time-to-event data, where the event of interest could be death, failure, recurrence of a disease, or any other event.

The Kaplan-Meier estimator calculates the probability of surviving beyond a given time point based on observed survival times in a sample. It considers censoring information, which indicates whether an individual's survival time is observed completely or only partially.

The estimator's mathematical formula for the survival probability at each time point t is given by:

$$S(t) = S(t-1) * (1 - (d(t) / n(t)))$$

Where:

S(t) represents the estimated survival probability at time t.

S(t-1) represents the estimated survival probability at the previous time point.

d(t) represents the number of events (deaths or failures) at time t.

n(t) represents the number of individuals at risk just before time t.

By calculating the survival probabilities at different time points, the Kaplan-Meier estimator generates a survival curve that depicts the estimated survival function over time. The survival curve visually represents the probability of surviving beyond each time point and provides valuable insights into the survival patterns of a population.

The Cox Proportional Hazards model:

The Cox Proportional Hazards model, also known as Cox regression, is a widely used statistical model in survival analysis. It allows us to investigate the relationship between the predictor variables (covariates) and the hazard function, which represents the risk of an event occurring over time. The model is particularly useful for analyzing censored time-to-event data.

The Cox Proportional Hazards model assumes that the hazard function for an individual is a product of two components: a baseline hazard function that describes the hazard when all covariates are zero, and a set of covariate effects that modify the hazard for everyone. The model does not make any assumptions about the shape of the baseline hazard or the functional form of the covariate effects.

$$h(t|X) = h0(t) * \exp(b1X1 + b2X2 + ... + bp*Xp)$$

h(t|X) represents the hazard rate at time t for an individual with a set of covariate values X.

h0(t) is the baseline hazard function, which represents the hazard when all covariates are zero. It describes the underlying risk of the event over time.

exp(b1X1 + b2X2 + ... + bp*Xp) represents the exponential term, which is the product of the regression coefficients (b1, b2, ..., bp) and the corresponding covariate values (X1, X2, ..., Xp). This term modifies the baseline hazard based on the covariate effects.

```
# Kaplan Meier Method
kmf = KaplanMeierFitter()
kmf.fit(df['tenure'], event_observed=df['Churn'])
# Cox Proportional Hazards model
cph = CoxPHFitter()
cph.fit(df, duration_col='tenure', event_col='Churn')
```

Results of the models:

The Results of both these models are then plotted in the Fig 2 and the Cox proportional hazard has lesser RMSE(Root Mean Square error) and MAE (Mean Average Error) than Kaplan Meier meaning that the Cox method is much closer in prediction as shown in the table below.

Model	RMSE (Root Mean Square error)	MAE (Mean Average Error)
Kaplan Meier	0.49387630886567196	0.4079094948125434
Cox proportional hazard Model	0.4344650497476844	0.36843875874248694

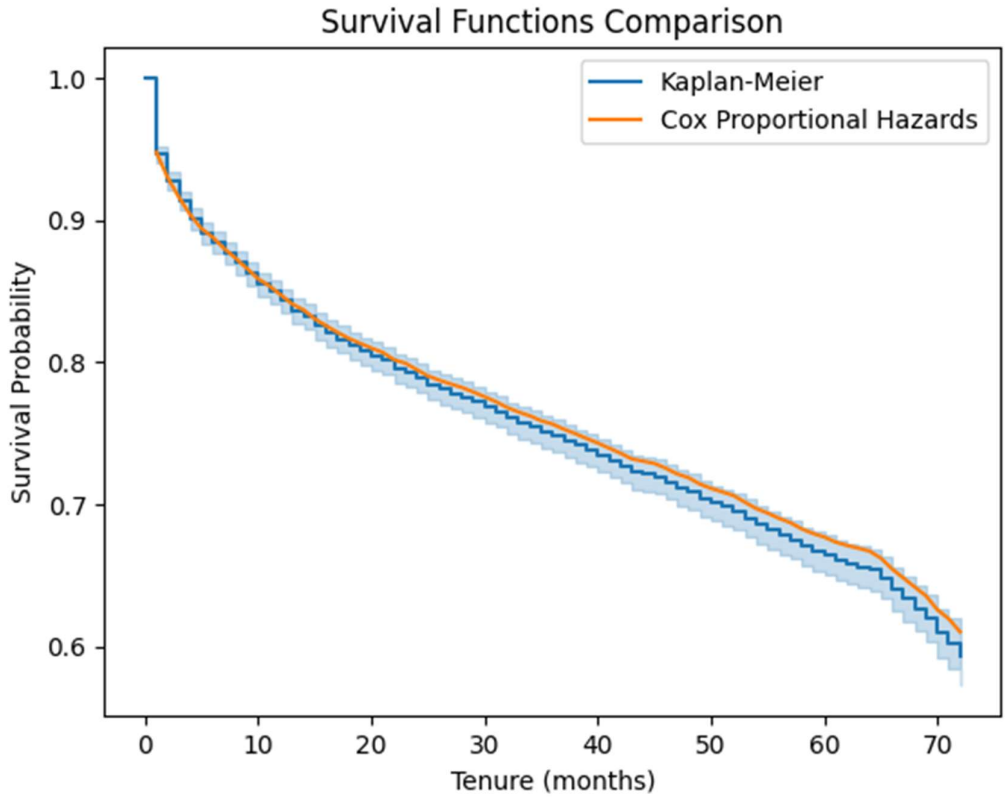

Figure 2: Survival analysis of kaplan Meier and Cox proportional hazard model

When dealing with these 2 models I have used only the RMSE and MSA to determine their performance and from looking at the values predicted I would argue that Cox hazard model is better at predicting the Users survival rate as it has lesser RMSE where the average between the predicted and actual value is higher it will penalize it a lot more than MAE , but since both the factors favour COX model , I would conclude that Cox is superior for this dataset used.

Other models and their results:

Since we have looked at the two most used survival prediction models Kaplan Meier and Cox proportional hazard Model let's look at two more different models such as Logistic Regressionn and Support Vector Machine.

Logistic Regression:

Logistic Regression is a statistical modeling technique used to predict the probability of a binary outcome based on a set of predictor variables. In survival analysis, logistic regression can be used to model the probability of survival (or the hazard) over time. It allows us to estimate the effects of different predictors on the probability of survival while considering censoring, which is a common issue in survival data where the event of interest has not occurred for all individuals.

Logistic regression assumes that the log odds of survival (logit) are a linear combination of predictor variables.

Support Vector Machine:

Support Vector Machine (SVM) is a machine learning algorithm that can be used for both classification and regression tasks. In survival analysis, SVM can be applied to estimate the survival function by finding a hyperplane that separates the survival times of different groups or predicts the hazard function. SVM aims to find an optimal decision boundary that maximally separates the survival times or minimizes the misclassification error. SVM can handle non-linear relationships between predictors and survival outcomes by using kernel functions.

```
models = {
    'Logistic Regression': make_pipeline(preprocessing,
LogisticRegression(max_iter=1000)),
    'Support Vector Machine': make_pipeline(preprocessing,
SVC(probability=True))
}

# Train and evaluate each model
for name, model in models.items():
    print(f"Training {name}...")
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    y_pred_proba = model.predict_proba(X_test)[:, 1]  # Probability of
positive class (churn)

    rmse = mean_squared_error(y_test, y_pred) ** 0.5
    mae = mean_absolute_error(y_test, y_pred)

    print(f"RMSE: {rmse}")
    print(f"MAE: {mae}")
    print("\n")
```

Discussion on the above models:

Both the above models have their own strengths and weaknesses. But both these models have better RMSE and MAE scores than the Kaplan Meier and Cox proportional hazard Model as shown in the below table.

Model	RMSE (Root Mean Square error)	MAE (Mean Average Error)
Logistic Regression	0.4245817881386584	0.18026969481902058
Support Vector Machine	0.44498626385287793	0.1980127750177431

The issue is that these metrics are not what should be used as we evaluate survival-based metrics for a better model . We need to use the C-index and Log-Rank Test.

Concordance Index (C-index):

The C-index measures the discriminatory power of the model. It indicates the probability that, given two randomly selected individuals, the model correctly predicts the order of their survival

times. A value of 0.5 indicates a random prediction, while a value of 1.0 represents a perfect prediction.

Log-Rank Test:

This is a statistical test used to compare survival curves between different groups or covariate levels. It assesses whether there is a significant difference in survival between the groups under investigation.

These 2 metrics are the most important factors in determining the better survival model. Hence, we can look at the C-index for both Kaplan Meier and Cox proportional hazard Model in the table below.

Model	C-index
Kaplan Meier	0.9888254298936076
Cox proportional hazard Model	0.5511228855420562

As Kaplan Meier has the higher c-index we can say that it is the better model overall for a survival analysis model. Thus, I will conclude that the Kaplan Meier is the better model for a survival analysis model as it has a higher C-index than Cox and other models.

Conclusion:

The Above we have taken the telco Dataset and have tested it against multiple models such as Kaplan Meier and Cox proportional hazard Model as they are traditionally used in survival analysis. But we also noticed that the models have a lower RMSE and MAE values when compared to logarithmic regression and support vector machines. This is then understood that the models are not only determined by the accuracy of prediction but also the Concordance Index which measure that when random events are taken the prediction is correct or not which is more relevant to survival analysis and hence, we have concluded that Kaplan Meier model is the better one when compared against the above dataset.

References:

1. Edward L. Kaplan and Paul Meier. “Nonparametric Estimation from Incomplete Observations”. In: Journal of the American Statistical Association 53.282 (1958), pp. 457–481. DOI: 10.1080/01621459.1958.10501452.

2. David R. Cox. “Regression Models and Life-Tables”. In: Journal of the Royal Statistical Society: Series B (Methodological) 34.2 (1972), pp. 187–202. DOI: 10.1111/j.2517-6161.1972.tb00899.x.

3. Jerald F. Lawless. Statistical Models and Methods for Lifetime Data. Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons, 1982.

4. Telco Customer Churn. Dataset. 2017. URL: https://www.kaggle.com/blastchar/telco-customerchurn.