DOMINIKA MATUSIAK

UNPOPULAR ANALYSIS OF POPULAR DATA SET

FEBRUARY 2024

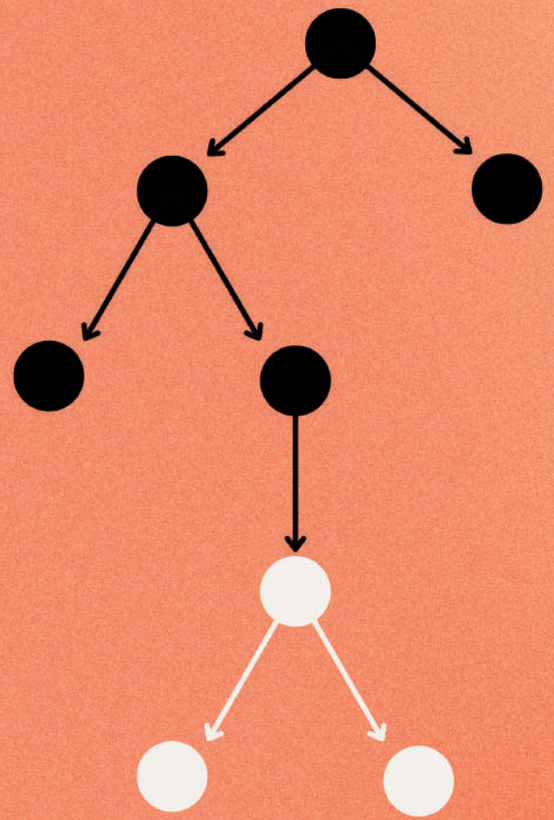
MATUSIAKDK@GMAIL.COM

CHURN ANALYSIS

WHEN DO CUSTOMERS CHURN?

WHAT ARE THE CONTRIBUTING FACTORS?

Churn	AccountWeeks	ContractRenewal	DataPlan	CustServCalls	DayCalls	MonthlyCharge	OverageFee	RoamMins	DataUsage_Group	DayMins_Group
0	128	1	1	1	110	89.0	9.87	10.0	4	4
0	107	1	1	1	123	82.0	9.78	13.7	4	2
0	137	1	0	0	114	52.0	6.06	12.2	1	4
0	84	0	0	2	71	57.0	3.1	6.6	1	4
0	75	0	0	3	113	41.0	7.42	10.1	1	2
0	118	0	0	0	98	57.0	11.03	6.3	1	4
0	121	1	1	3	88	87.3	17.43	7.5	4	4



POPULAR CHURN ANALYSIS

SURVIVAL ANALYSIS THEORY

1 Decision tree

\+ Intuitive and Easy to Interpret: splits data based on feature values

Handles Non-Linear Relationships

\- Prone to overfitting and can create overly complex models

Instability; small changes can lead to a completely different tree

2 Random forest

\+ High Accuracy by utilizing multiple decision trees making it highly reliable

Offers insights into which features are most influential

\- Creates more complex model harder to interpret than a single decision tree.

Computationally intensive, which leads to longer training times.

3 Gradient boost

\+ Sequential Improvement leading to high accuracy in predictions.

Can optimize on loss functions and provides parameter tuning options

\- Longer to train compared to other algorithms that allow parallelisation

If not carefully tuned can overfit to the training data

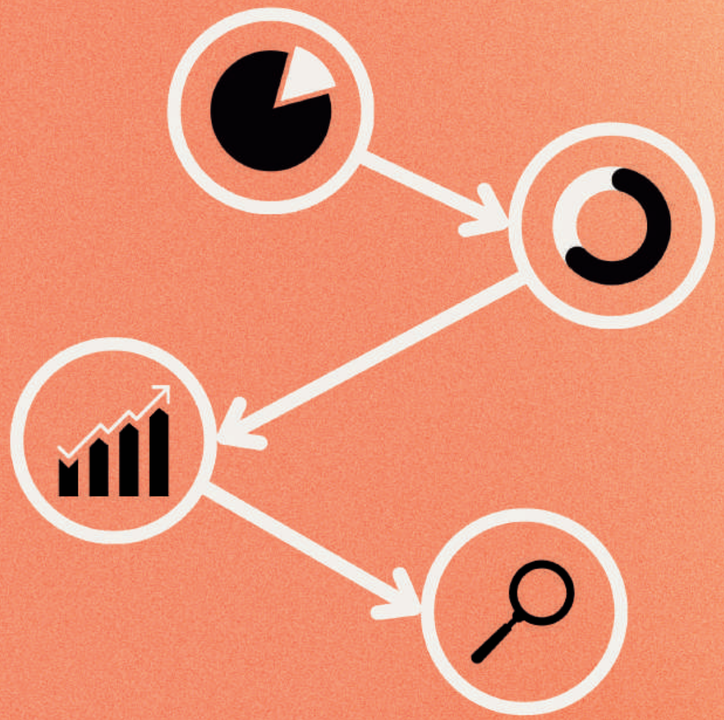
4 Knn

\+ Effectively classifies customers based on their similarity

No assumptions about data distribution making it versatile

\- Can be computationally expensive and slow

Sensitive to the scale of the data and irrelevant features



EDA

(EXPLORATORY DATA ANALYSIS)

a technique used in data analysis to summarize the main characteristics of a dataset, often with visual methods, to discover patterns, spot anomalies, and test hypotheses.

EXPLORATORY DATA ANALYSIS: STEPS

1 DATA COLLECTION → 2 DATA CLEANING → 3 DATA PROFILING

4 DATA VISUALIZATION → 5 DESCRIPTIVE STATISTICS → 6 CORRELATION ANALYSIS

7 DIMENSIONALITY REDUCTION → 8 FEATURE ENGINEERING → 9 HYPOTHESIS TESTING

WHAT DATA ARE WE WORKING WITH?

Attributes and their types

```
column_types = df.dtypes
print("Column Types:")
print(column_types)
```

```
Column Types:
Churn                int64
AccountWeeks         int64
ContractRenewal      int64
DataPlan             int64
DataUsage          float64
CustServCalls        int64
DayMins            float64
DayCalls             int64
MonthlyCharge      float64
OverageFee         float64
RoamMins           float64
dtype: object
```

Looking for NULL values

```
# Check for missing values in the dataset
missing_values = df.isnull().sum()

# Output the missing values
print("Missing values in each column:")
print(missing_values)

continuous_columns = ['DataUsage',
                      'DayMins',
                      'MonthlyCharge',
                      'OverageFee',
                      'RoamMins']
```

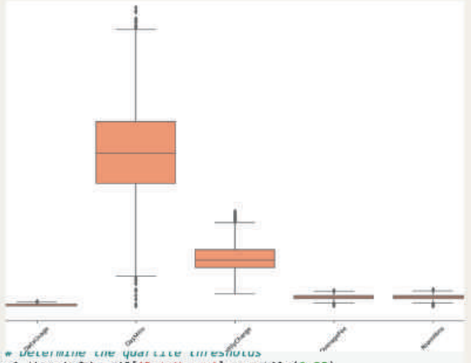
```
Missing values in each column:
Churn             0
AccountWeeks      0
ContractRenewal   0
DataPlan          0
DataUsage         0
CustServCalls     0
DayMins           0
DayCalls          0
MonthlyCharge     0
OverageFee        0
RoamMins          0
```

Summary statistics

```
continuous_columns = ['DataUsage',
                      'DayMins',
                      'MonthlyCharge',
                      'OverageFee',
                      'RoamMins']
continuous_summary_statistics = df[continuous_column
print("Summary Statistics for Continuous Variables:"
print(continuous_summary_statistics)
```

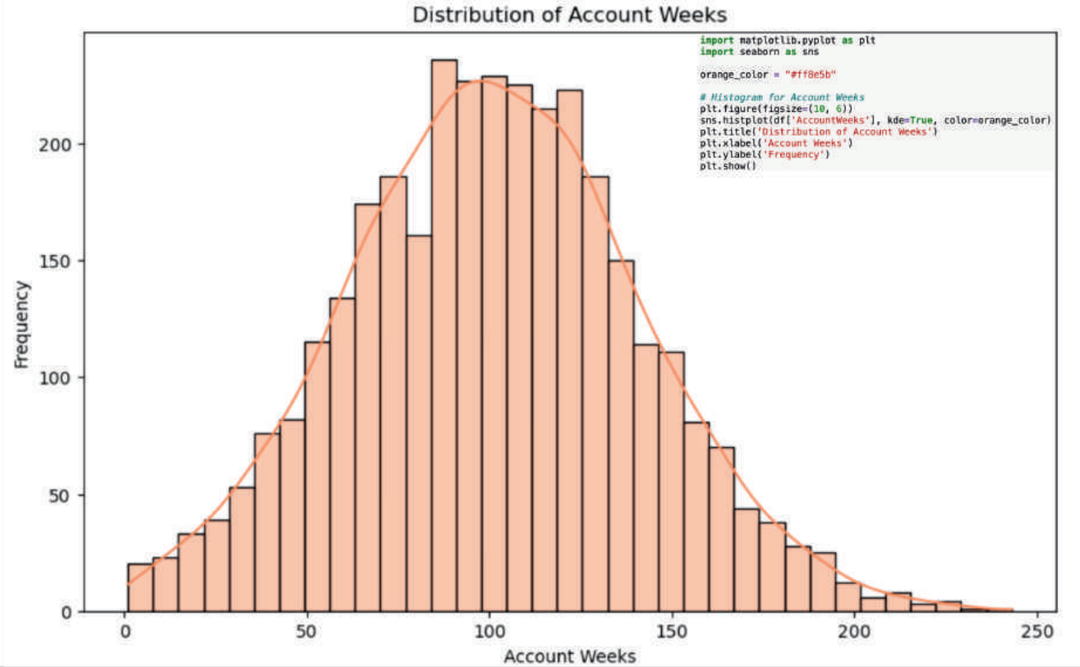
```
Summary Statistics:
              Churn  AccountWeeks  ContractRenewal
count  3333.000000   3333.000000      3333.000000  3
mean      0.144914    101.064806         0.903090
std       0.352067     39.822106         0.295879
min       0.000000      1.000000         0.000000
25%       0.000000     74.000000         1.000000
50%       0.000000    101.000000         1.000000
75%       0.000000    127.000000         1.000000
max       1.000000    243.000000         1.000000

          DataPlan    DataUsage
count  3333.000000  3333.000000
mean      0.276628     0.816475
std       0.447398     1.272668
min       0.000000     0.000000
25%       0.000000     0.000000
50%       0.000000     0.000000
75%       1.000000     1.780000
max       1.000000     5.400000
```

Looking for outliers

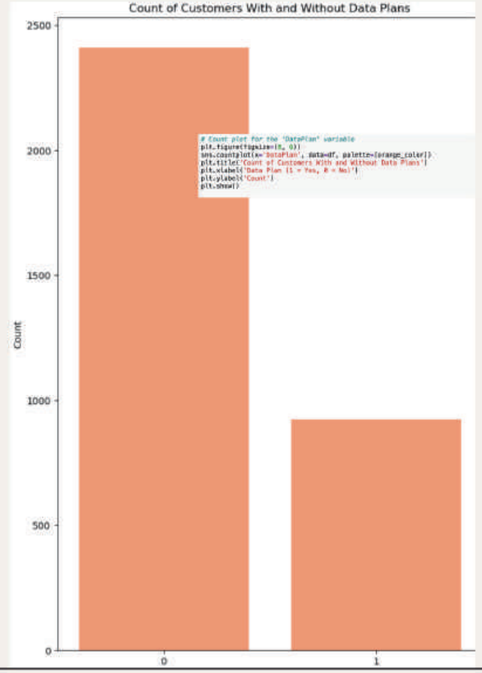


```
# Determine the quartile thresholds
q1_threshold = df['DataUsage'].quantile(0.25)
q2_threshold = df['DataUsage'].quantile(0.5) # median
q3_threshold = df['DataUsage'].quantile(0.75)

# Categorize 'DataUsage' into quartiles
df['DataUsage_Group'] = 4
df.loc[df['DataUsage'] <= q3_threshold, 'DataUsage_Group'] = 3
df.loc[df['DataUsage'] <= q2_threshold, 'DataUsage_Group'] = 2
df.loc[df['DataUsage'] <= q1_threshold, 'DataUsage_Group'] = 1

# Drop the original 'DataUsage' variable
df = df.drop('DataUsage', axis=1)
```

Histogram - distribution of account weeks



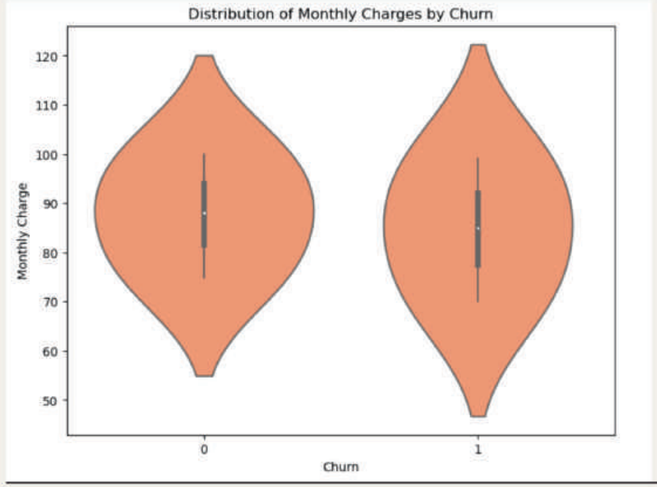
Count Plot - Data Plan and categorised Service Calls



THE NUMBER OF ACCOUNTS THAT DID NO CHURNED ARE 2.5 TIMES MORE !

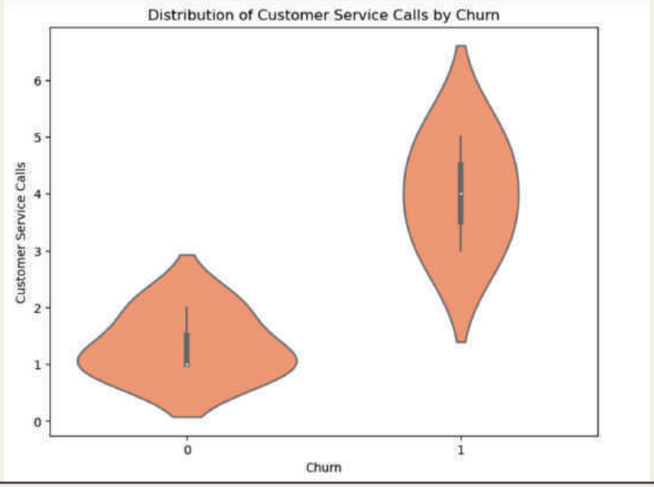
Monthly Charges by Churn

Similar distributions indicating **no significant variance in charges** that correlates with churn.



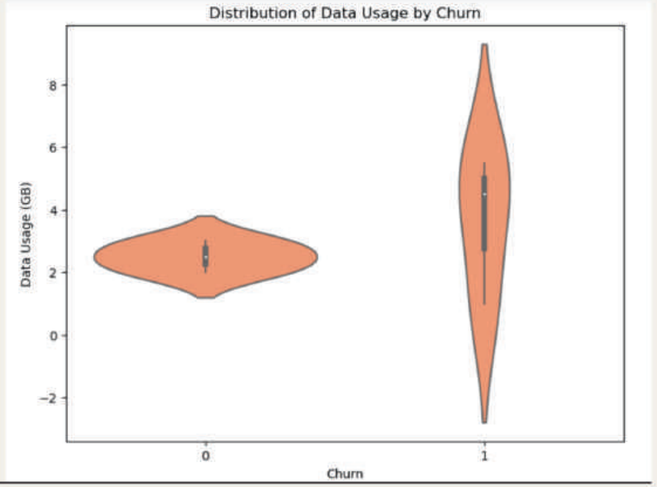
Customer Service Calls by Churn

Wider interquartile range suggesting **higher volatility in support engagement prior to churn.**



Data Usage by Churn

Greater median data usage with a wider variance for churned, **implying a correlation between increased data consumption and the likelihood of churn.**



MEANS COMPARISON OF TWO GROUPS

T-test for Customer Service Calls:

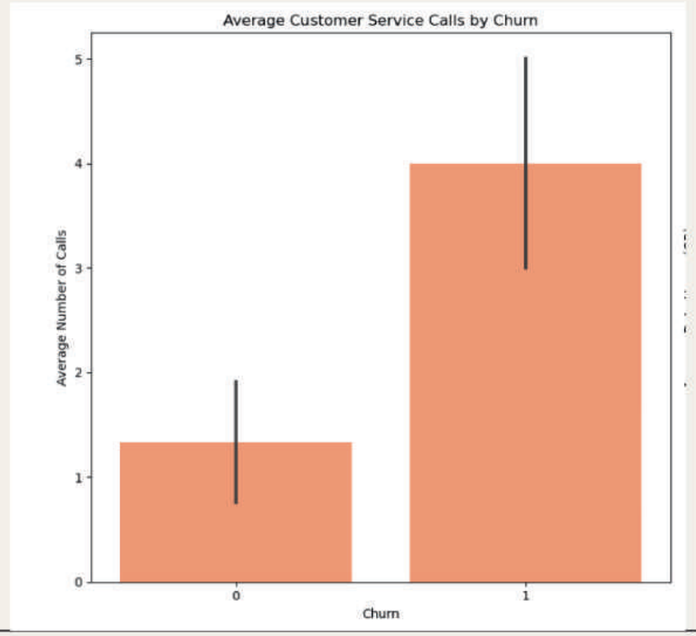
- T-statistic: 4.000000000000001
- P-value: 0.016130089900092518

T-test for Data Usage:

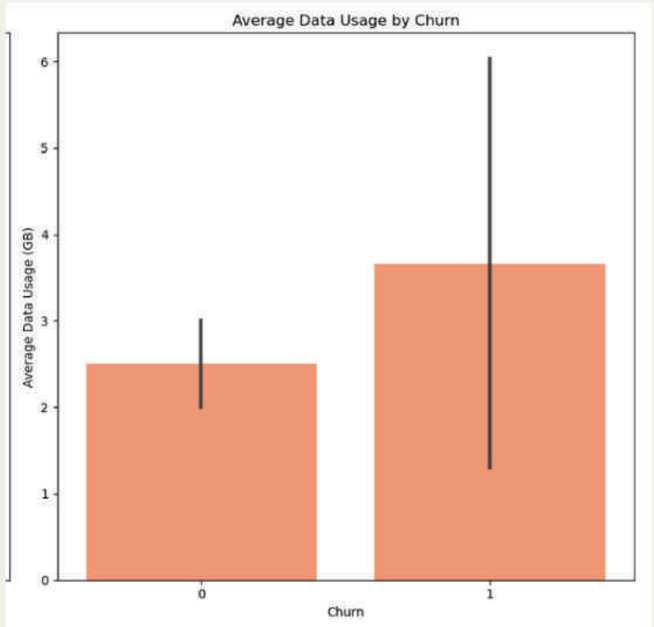
- T-statistic: 0.8366600265340753
- P-value: 0.4498551702745395

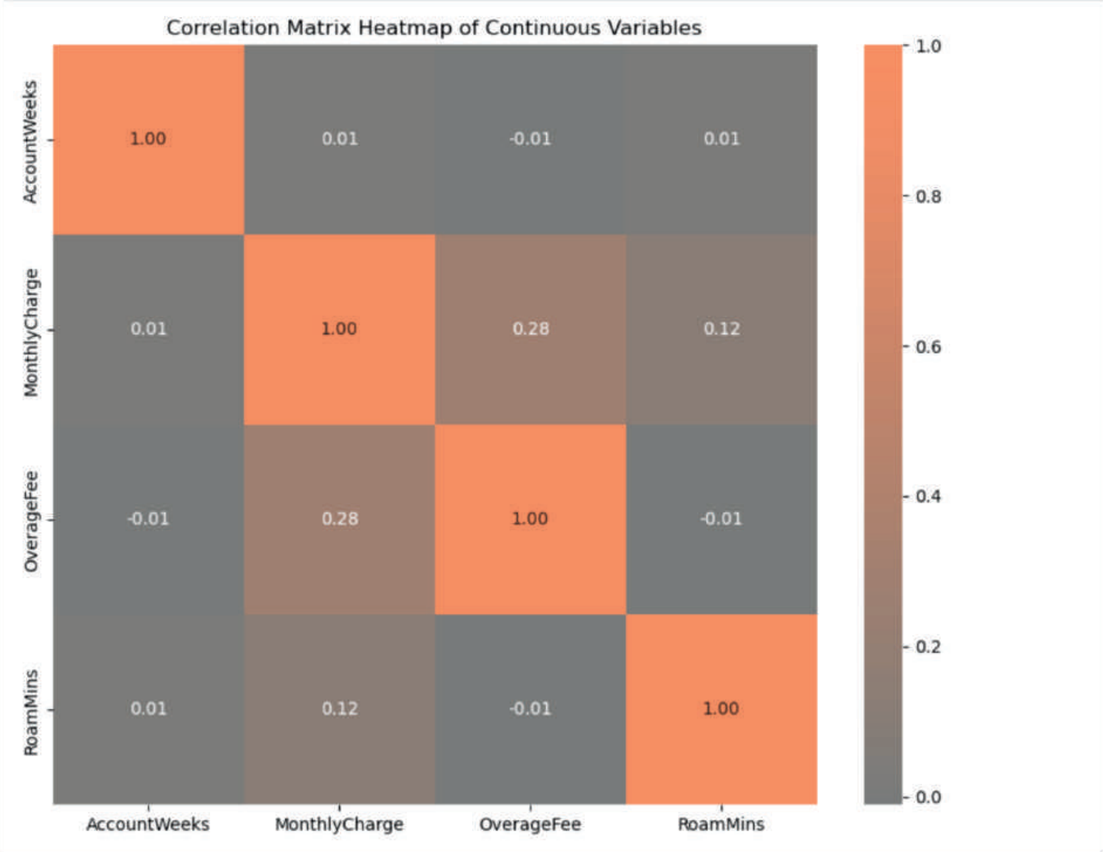
The T-test reveals that **churned customers make more service calls**, indicated by a positive T-statistic, but shows **no significant difference** in data usage **between churned and retained customers**.

Customer Service Calls by Churn T-test



Data Usage by Churn T-test



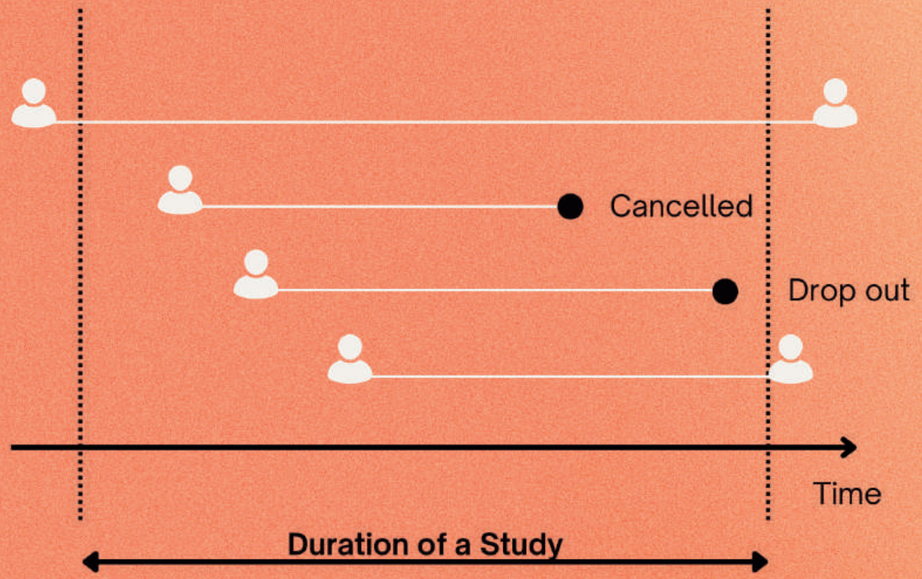


Pearson correlation coefficients in Python

- **Low R2 (coefficient of determination) -** only **7.84% (0.0784)** of the variability in OverageFee is accounted for by MonthlyCharge.

So what models can we use?

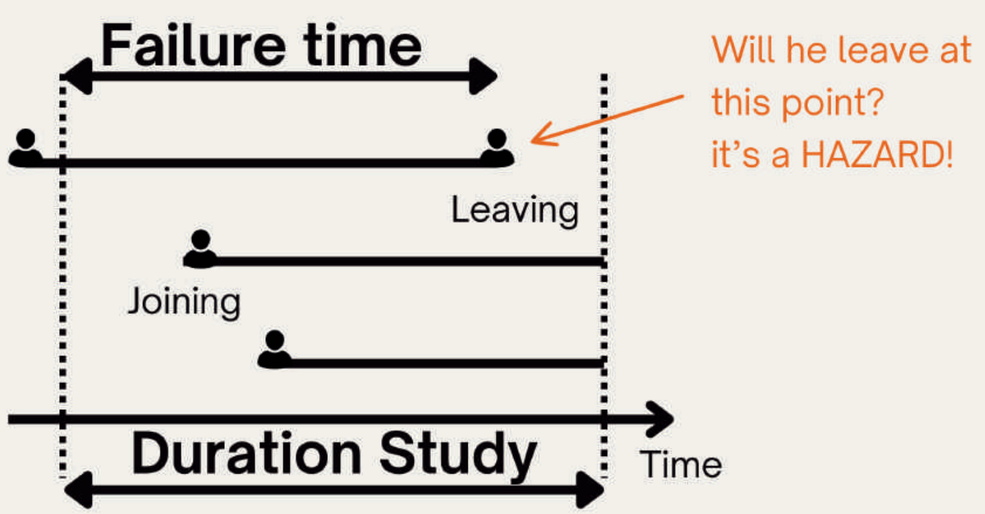
```
# Define the colors for uncorrelated and correlated
colors = ["#7b7c7a", "#ff8e5b"]
n_bins = 100
cmap = LinearSegmentedColormap.from_list("custom_corr", colors, N=n_bins)

plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, fmt='.2f', cmap=cmap, cbar=True)
plt.title('Correlation Matrix Heatmap of Continuous Variables')
plt.show()
```

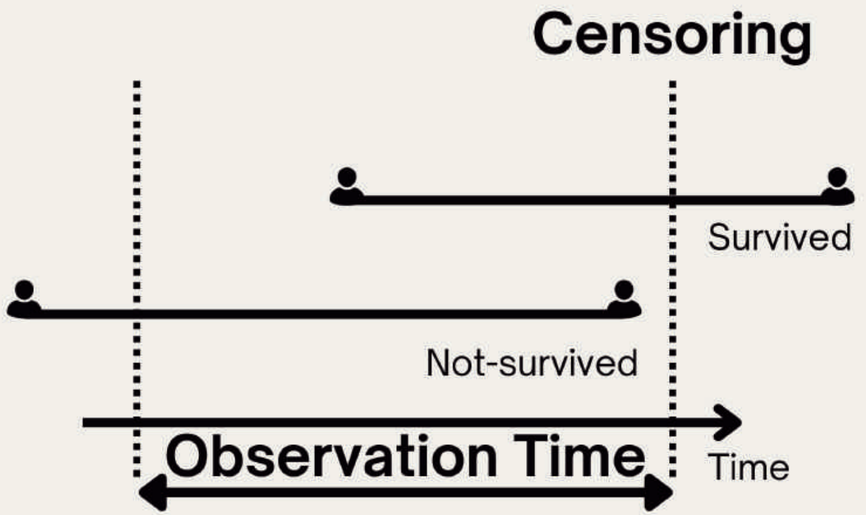


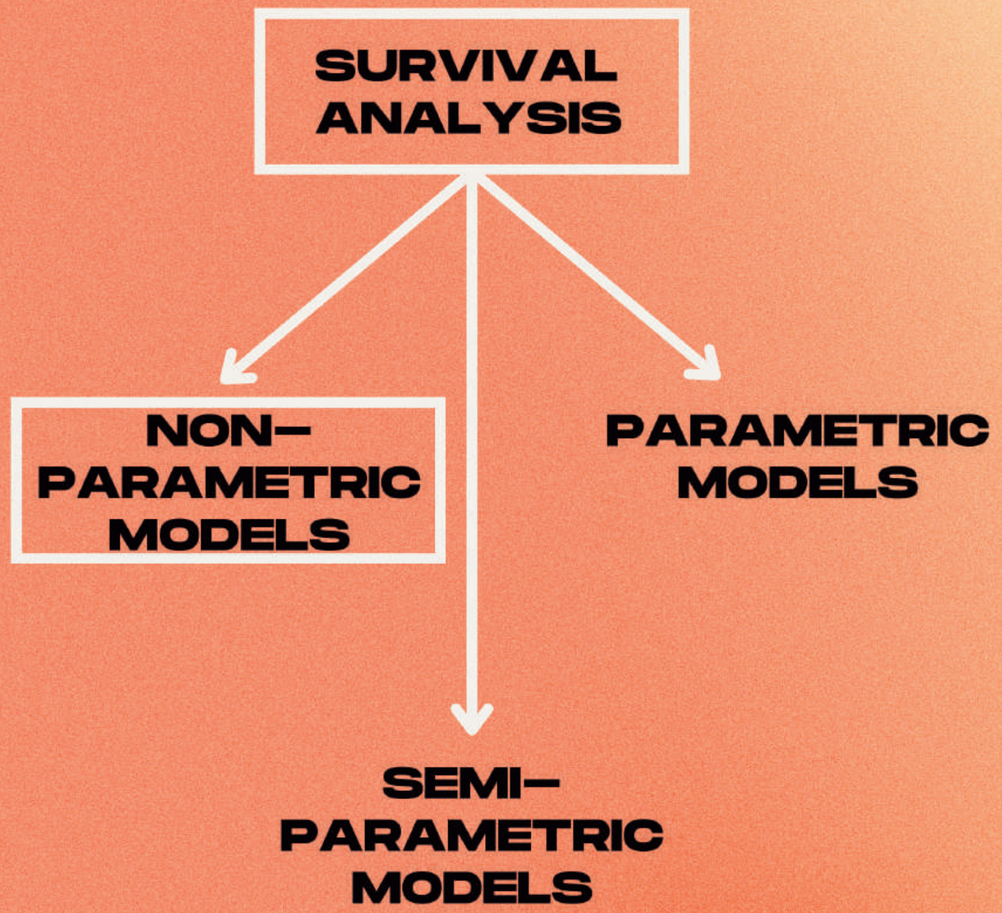
SURVIVAL ANALYSIS

a statistical approach used to determine the time until an event of interest occurs, such as customer churn. It accommodating situations where not all events are observed within the study period (censoring).

What are the key concepts?







WHAT METHODS CAN WE USE

NON-PARAMETRIC MODELS

WHEN TO USE?

Used for **initial data exploration without** imposing rigid **parametric assumptions** on the data.

\+

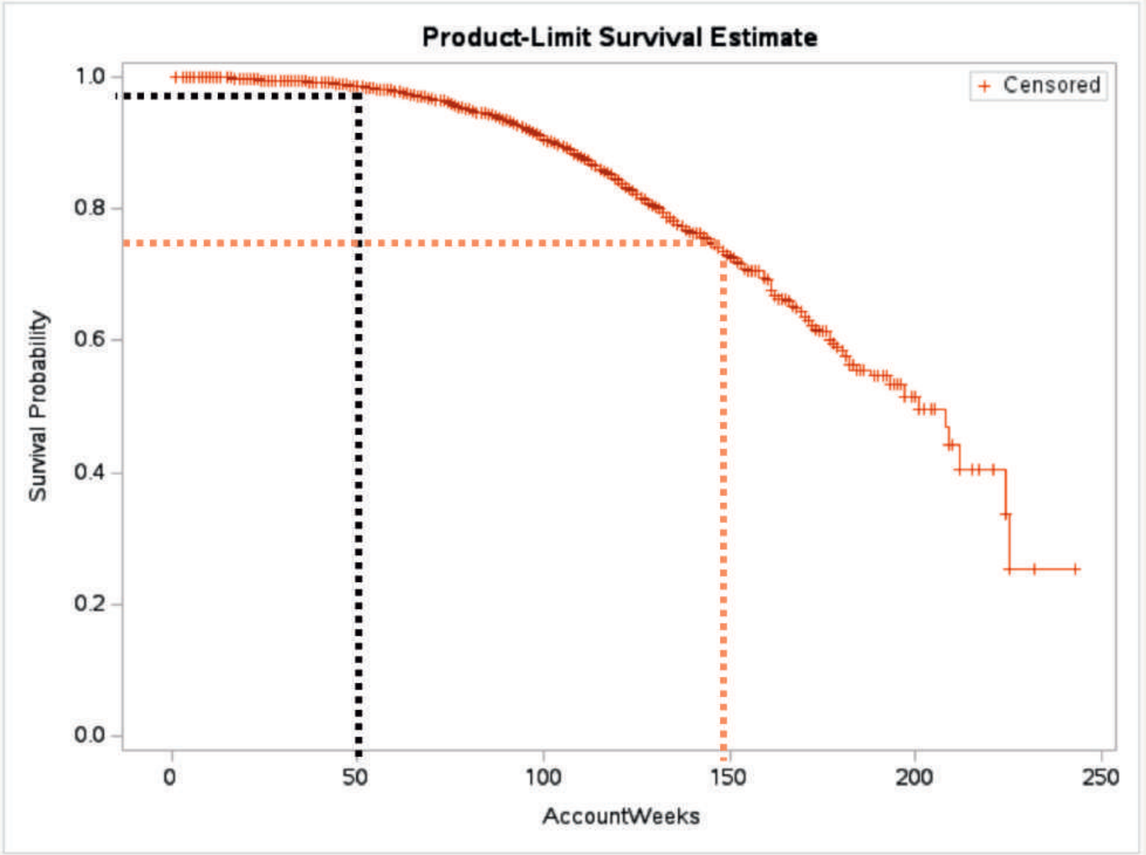
- **Not requiring assuming a specific distribution** of survival times.
- Results, such as survival function estimates, are **easy to interpret.**
- They do not require complex calculations.

\-

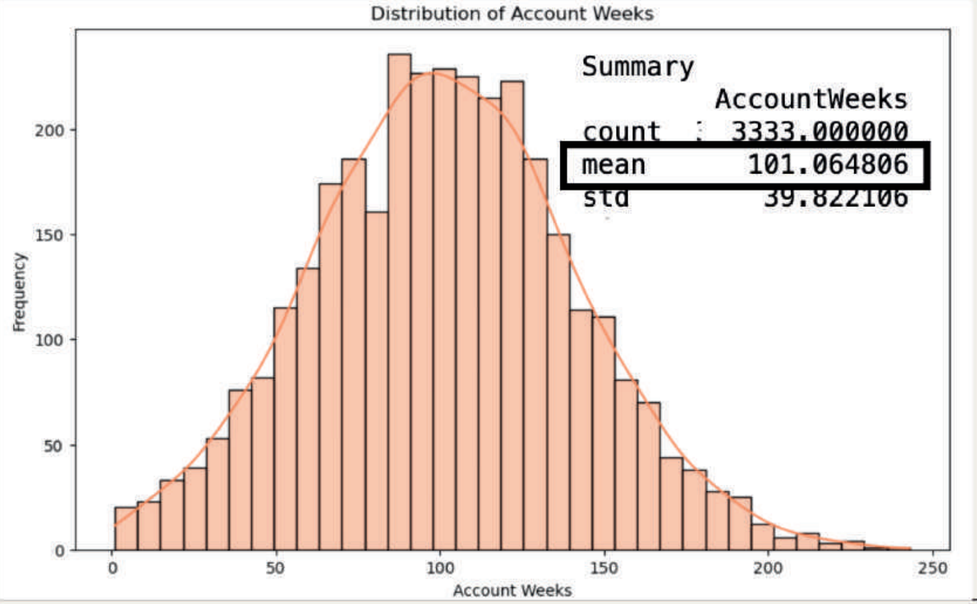
- They **cannot simultaneously incorporate many explanatory variables**.
- There can be interpretative **challenges with small samples or large datasets.**

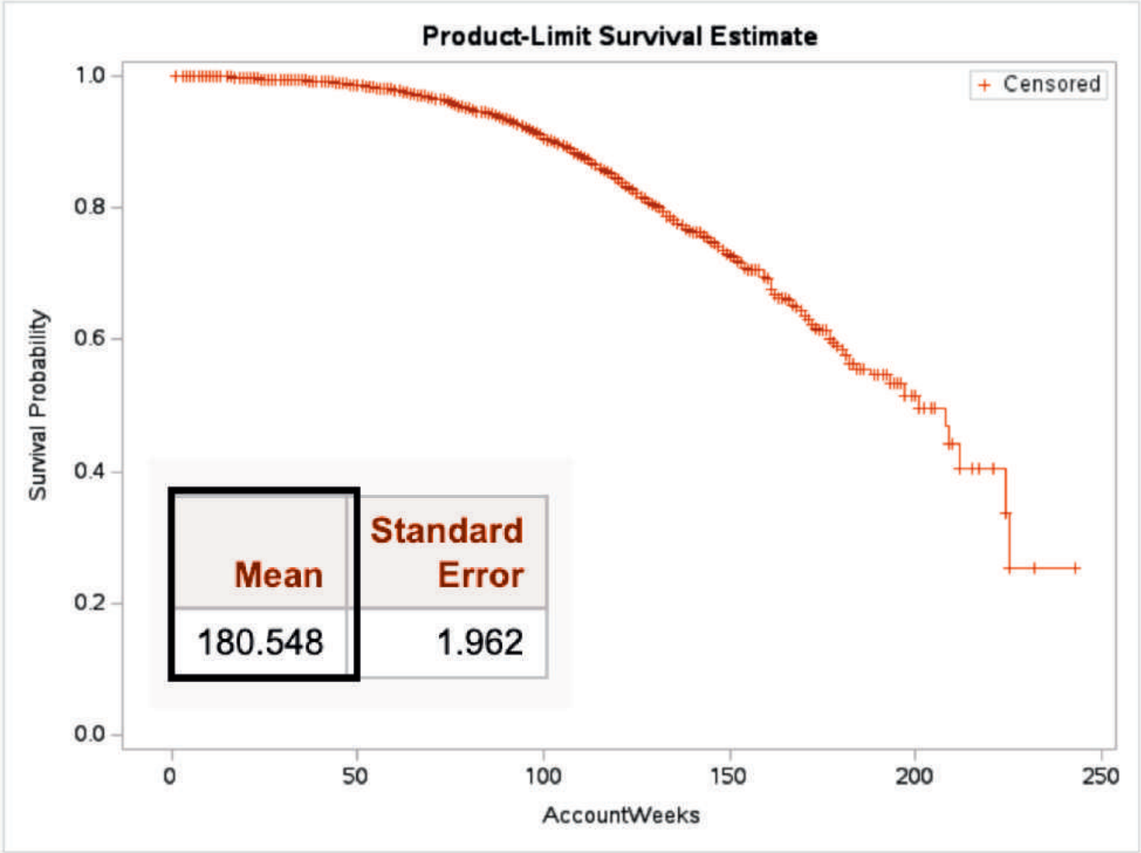
Non-parametric models - survival graphs and intervals

Interval		Number Failed	Survival	Failure
[Lower,	Upper)			
0	50	42	1.0000	0
50	100	185	0.9868	0.0132
100	150	200	0.9123	0.0877
150	200	50	0.7517	0.2483
200	250	6	0.5824	0.4176
250	.	0	0.3706	0.6294

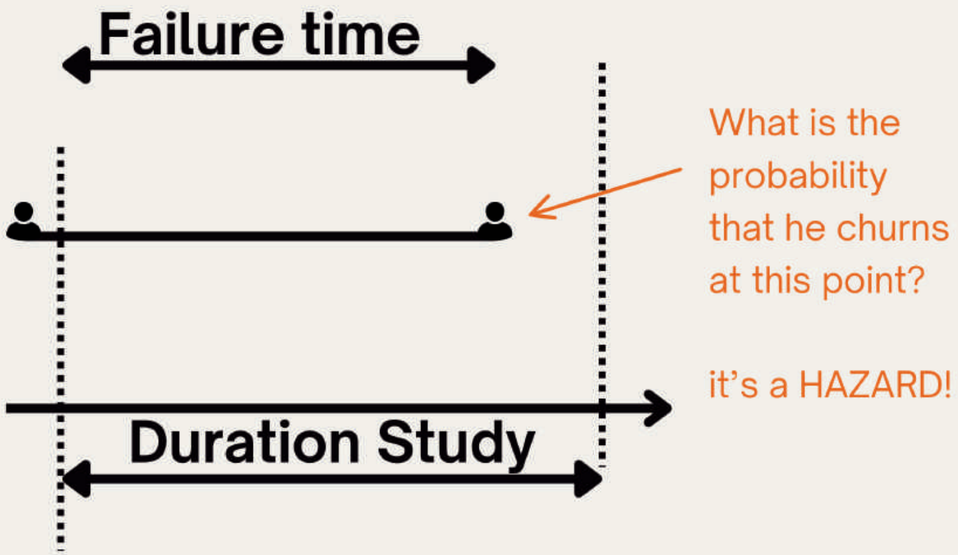


Histogram - distribution of account weeks vs non-parametric models



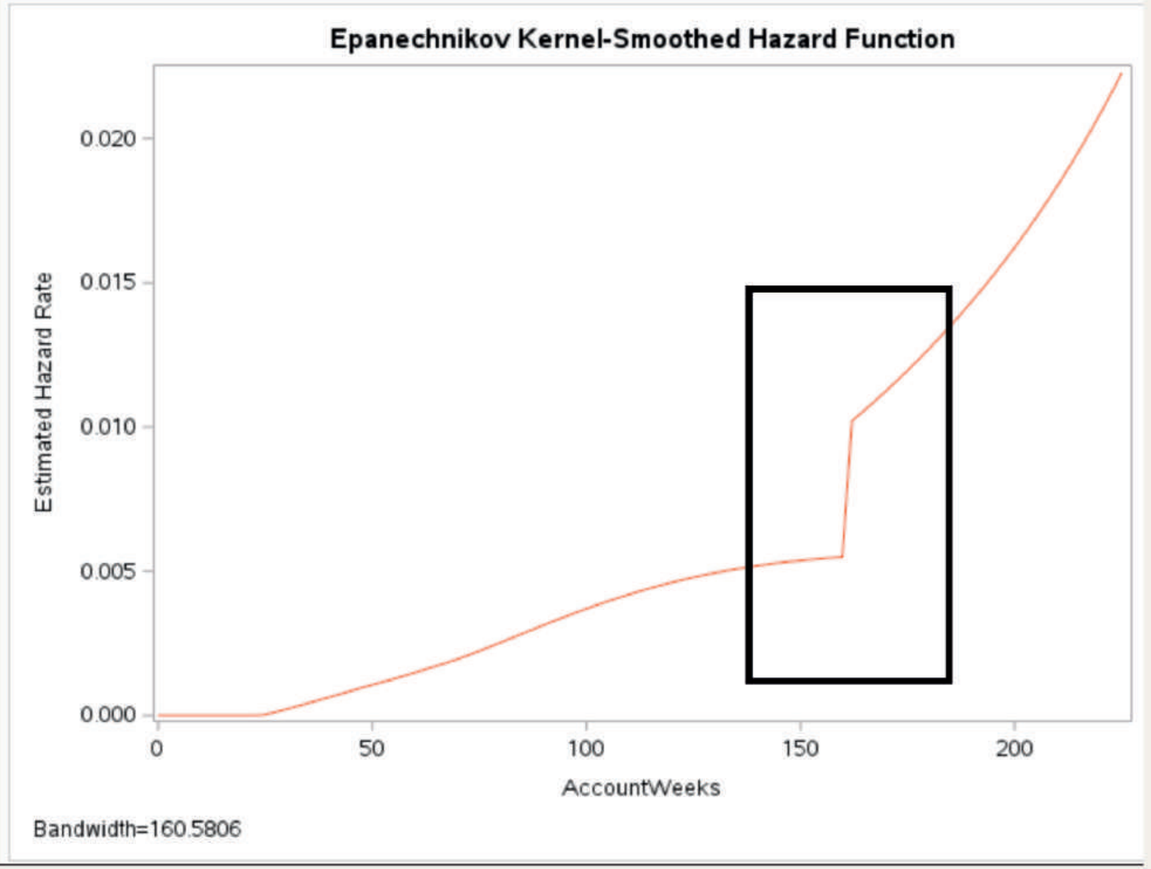


HAZARD FUNCTION



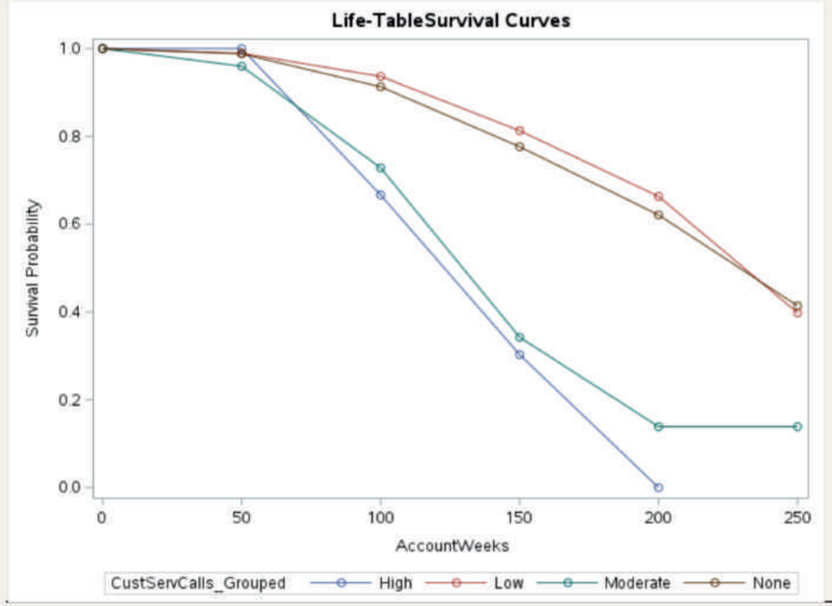
What is the probability that he churns at this point?

it's a HAZARD!



NON-PARAMETRIC MODELS: COMPARISON BETWEEN SUMMARY STATISTICS AND SURVIVAL MODEL

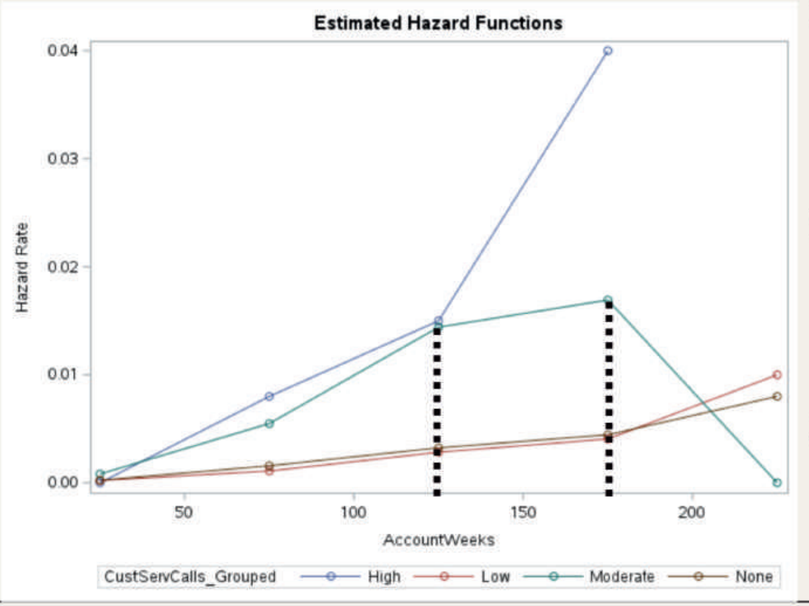
Survival function based on Customer Service Calls



Similar Probabilities of survival for None and Low

Similar Probabilities of survival for Moderate and High

Hazard function based on Customer Service Calls



Low Calls

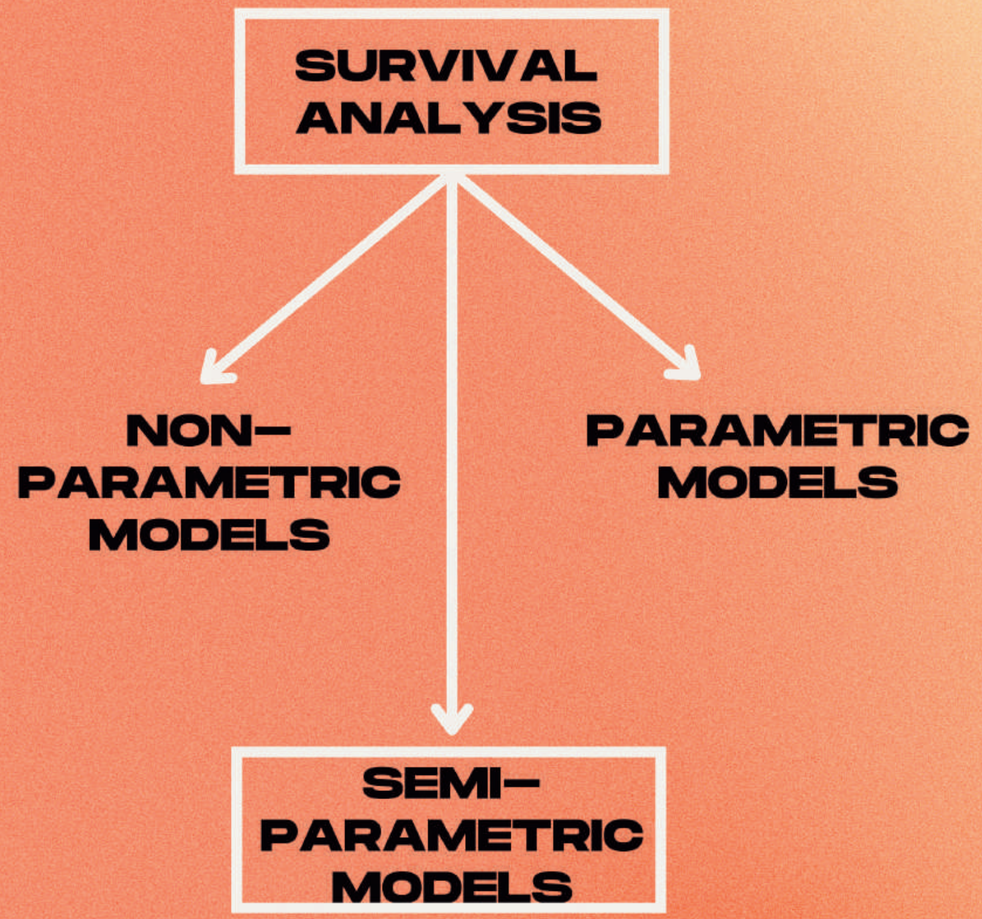
- Risk slightly increases around 175th week

Moderate Calls

- Risk decreases around 175th week

High Calls

- Risk increases around 125th week


SURVIVAL ANALYSIS
NON–PARAMETRIC MODELS
PARAMETRIC MODELS
SEMI–PARAMETRIC MODELS

NEXT STEP

SEMI-PARAMETRIC MODELS

WHEN TO USE?

when we want to **understand** the **contributing factors** to churn rate without assuming the survival distribution.

We assume that the **risk between subjects** (hazard ratio) is **constant over time**

\+

- Adaptable to different types of data.
- Can provide reliable estimates even when the hazard function's form is unknown or complex.
- The coefficients can be interpreted as effects on the hazard rate

\-

- The Cox model assumes that the ratio of hazard rates between any two individuals is constant over time - not always hold true.

SEMI-PARAMETRIC MODEL WITH MULTIPLE EXPLANATORY VARIABLE.

HOW TO SELECT CONTRIBUTING FACTORS?

Stepwise Procedure:

forward addition and backward variable removal

Step	Effect		
	Entered	Removed	Pr > ChiSq
1	CustServCalls_Groupe		<.0001
2	DayMins_Group		<.0001
3	RoamMins		<.0001
4	ContractRenewal		<.0001
5		RoamMins	0.1900

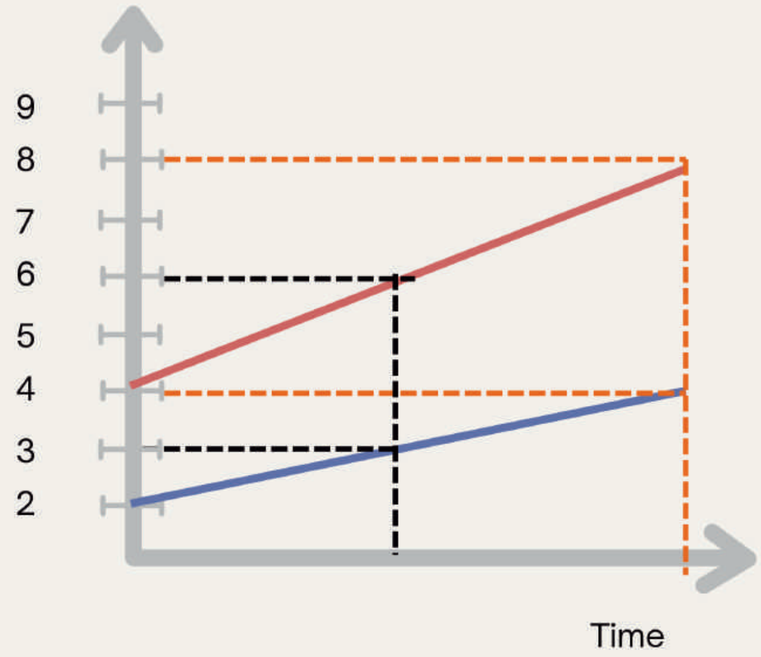
```
proc phreg data=dane.df_churn;
    class ContractRenewal DataPlan DayCalls RoamMins DataUsage_group
    DayMins_group CustServCalls_Grouped;
    model AccountWeeks*churn(0)=ContractRenewal DataPlan DayCalls RoamMins
    DataUsage_group DayMins_group CustServCalls_Grouped
/ties=efron selection=stepwise ;
run;
```

Higher hazard for non-contracted and moderate-frequency call customers

Consumers with a high number of daily minutes have a hazard rate that is 5 times lower compared to customers with a low number of daily minutes.

Parameter		Pr > ChiSq	Hazard Ratio	Label
ContractRenewal	0	<.0001	2.996	ContractRenewal 0
DayMins_Group	1	<.0001	0.450	DayMins_Group 1
DayMins_Group	2	<.0001	0.386	DayMins_Group 2
DayMins_Group	3	<.0001	0.230	DayMins_Group 3
CustServCalls_Groupe	High	0.0010	3.368	CustServCalls_Grouped High
CustServCalls_Groupe	Low	0.2475	0.868	CustServCalls_Grouped Low
CustServCalls_Groupe	Moderate	<.0001	4.074	CustServCalls_Grouped Moderate

WHAT IS HAZARD RATIO?



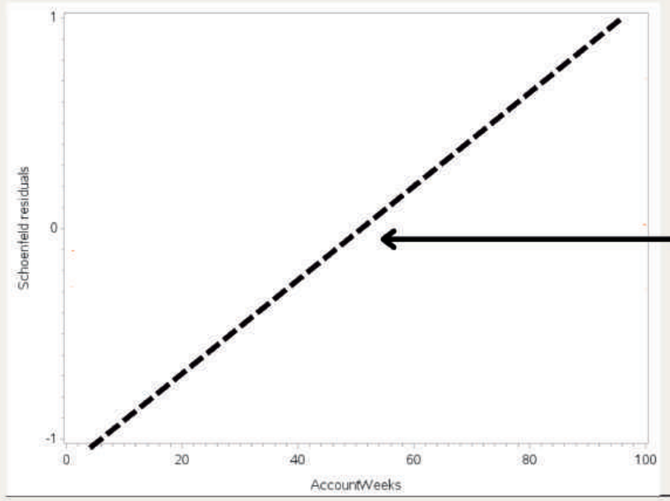
2 times higher risk through!

SEMI-PARAMETRIC MODEL WITH MULTIPLE EXPLANATORY VARIABLE.

HAZARD RATIOS ARE CONSTANT OVER TIME?

Yes - Schoenfeld Residuals

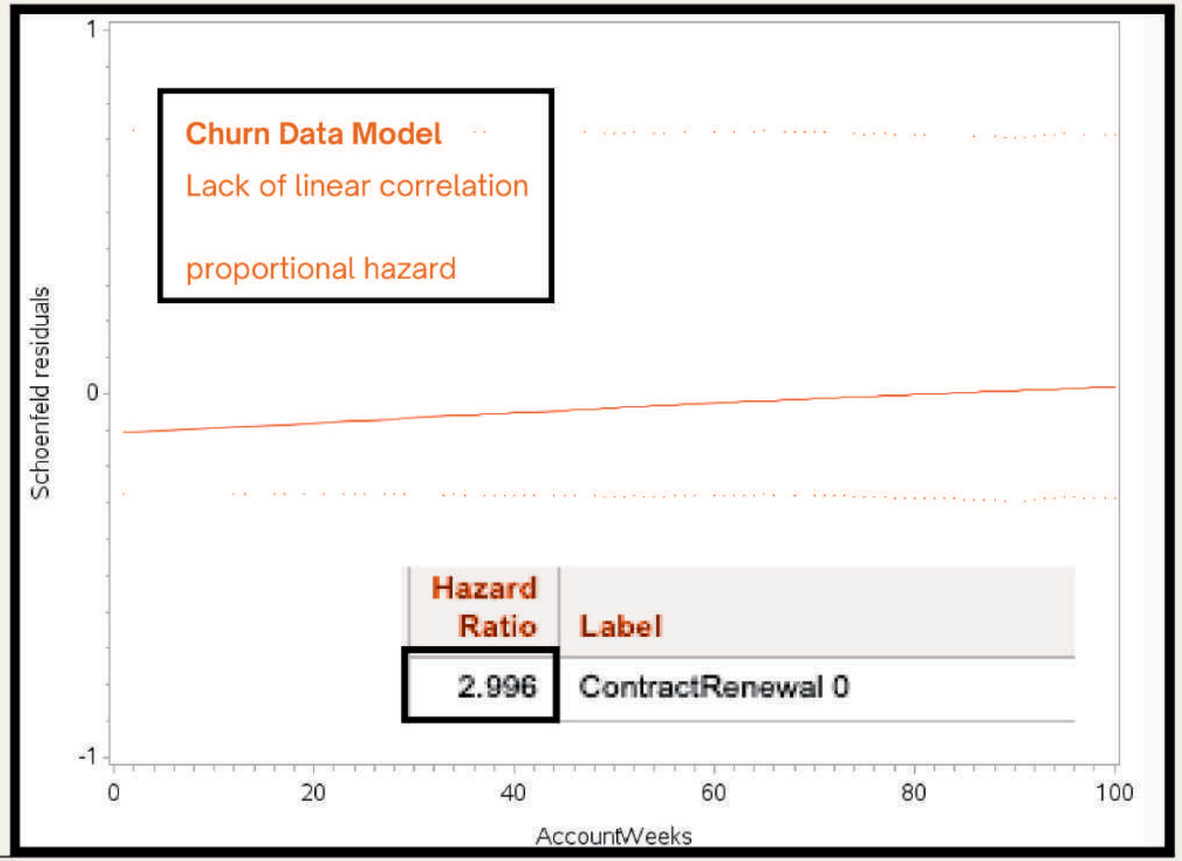
Lack of strong linear correlation for all variables



Example of a perfect linear correlation

values from -1 to 1

non-proportional hazard

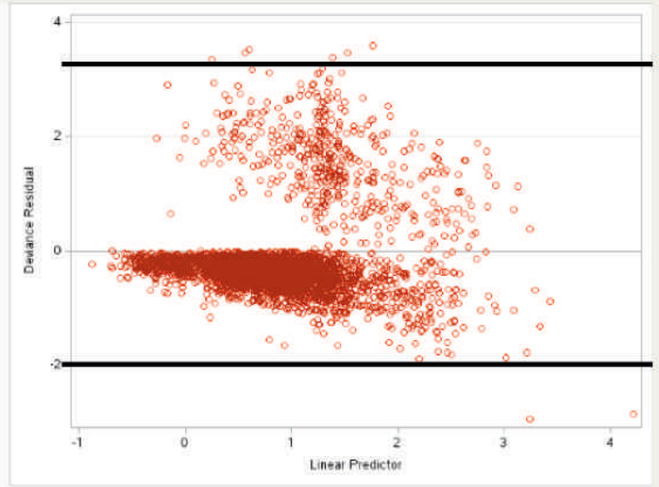


SEMI-PARAMETRIC MODEL WITH MULTIPLE EXPLANATORY VARIABLE.

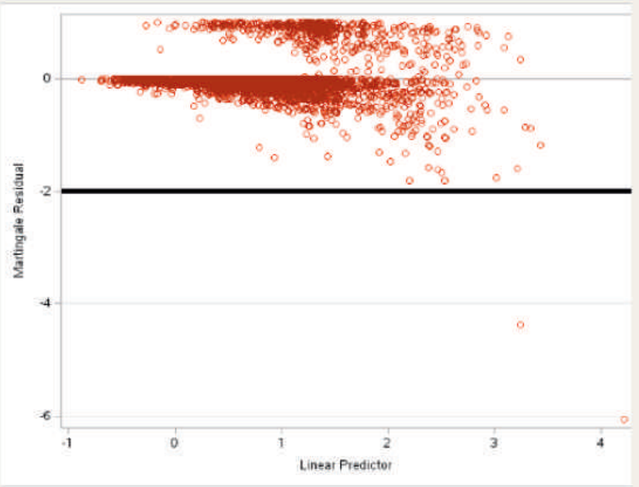
FINE-TUNING THE MODEL

good model fit and the absence of specification errors & outlier identification and removal.

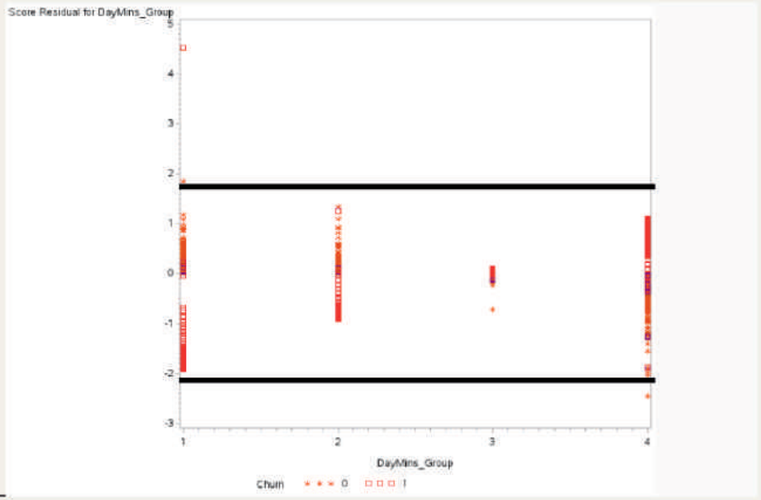
Martingale residuals



Deviance Residuals



Score Residuals



BEFORE AND AFTER – MODEL COMPARISON

Before cleaning

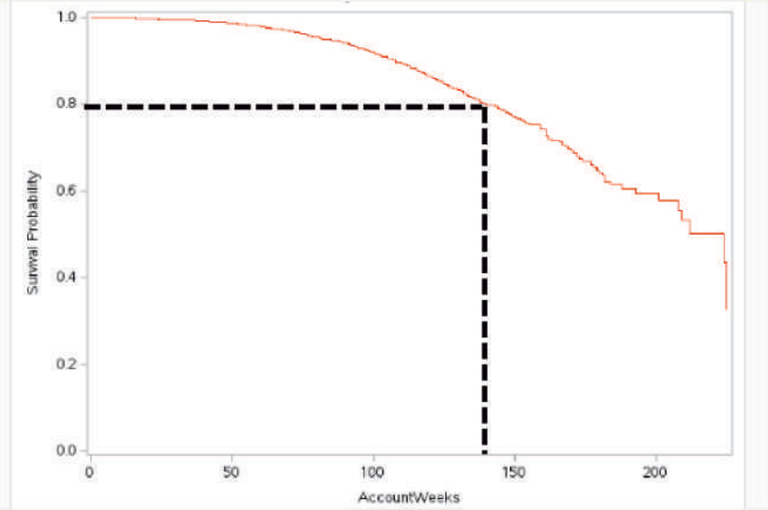
Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	6798.745	6135.207
AIC	6798.745	6471.207
SBC	6798.745	7172.052

After cleaning

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	6798.745	6323.944
AIC	6798.745	6337.944
SBC	6798.745	6367.146

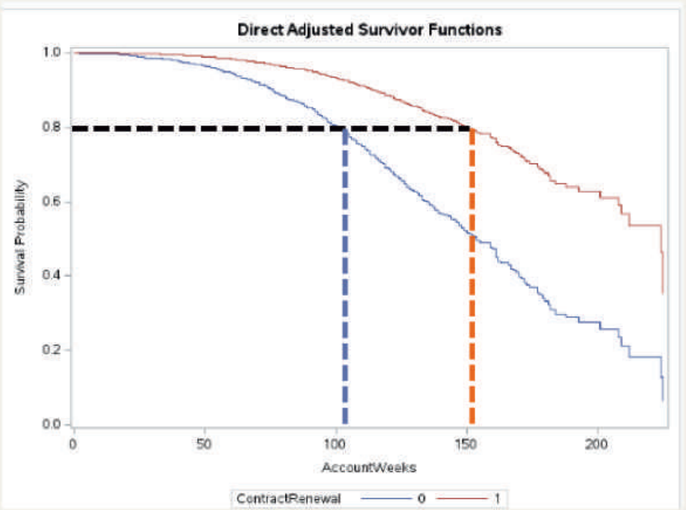
SEMI-PARAMETRIC MODEL: DIRECTLY ADJUSTED SURVIVAL FUNCTION

COMBINED VARIABLES



There is an 80% probability that a customer will remain with the company for at least 140 weeks.

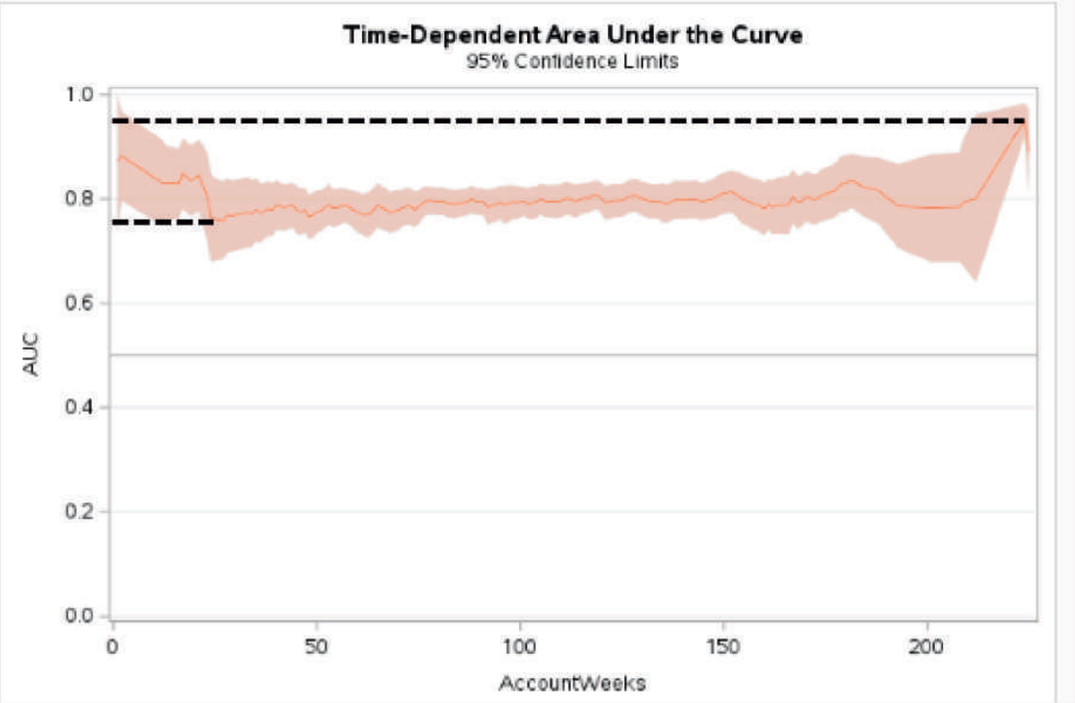
CONTRACT VS NO CONTRACT



There is an 80% probability that a customer with contract will remain with the company for at least 150 weeks and customer with no contract will remain with the company for at least 100 weeks.

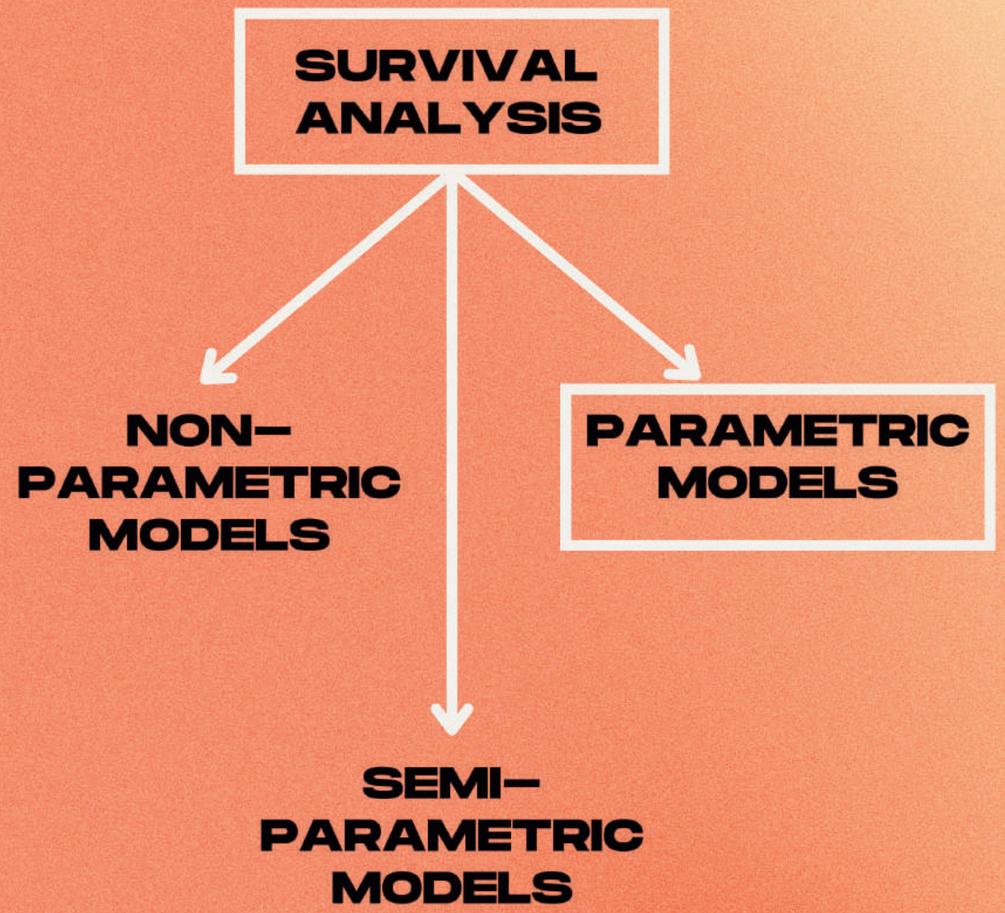
SEMI-PARAMETRIC MODEL: EVALUATION OF PREDICTION ACCURACY AT SPECIFIC MOMENTS IN TIME

BUT WHAT'S THE MODEL ACCURACY



Depending on the time the model accuracy ranges between 80% and 97%

```
proc phreg data=dane.df_churn_po_usunieciu_2 plots(overlay=individual)=roc rocoptions(at= 28 56 112 224);
    class CustServCalls_grouped_num DayMins_group ContractRenewal;
    model AccountWeeks*churn(0)= CustServCalls_grouped_num DayMins_group ContractRenewal / ties=efron;
run;
proc phreg data=dane.df_churn_po_usunieciu_2 plots=auc rocoptions(method=ipcw(cl seed=1234) iauc);
    class CustServCalls_grouped_num DayMins_group ContractRenewal;
    model AccountWeeks*churn(0)= CustServCalls_grouped_num DayMins_group ContractRenewal / ties=efron;
run;
```



FURTHER ANALYSIS

PARAMETRIC MODELS

WHEN TO USE?

when the survival time distribution is well-understood, especially if you need to conduct simulation studies

WHY IT IS NOT USED NOW

validity heavily depends on the correctness of the assumed distributional form; a wrong assumption can lead to biased results.

DOMINIKA MATUSIAK

THANK YOU

FEBRUARY 2024

MATUSIAKDK@GMAIL.COM