Summary/Review

**Deep Learning for Time Series Forecasting**

Neural networks offer several benefits over traditional time series forecasting models, including:

* Automatically learn how to incorporate series characteristics like trend, seasonality, and autocorrelation into predictions.
* Able to capture very complex patterns.
* Can simultaneously model many related series instead of treating each separately.

Some disadvantages of using Deep Learning for Time Series Forecasting are:

* Models can be complex and computationally expensive to build (GPUs can help).
* Deep Learning models often overfit.
* It is challenging to explain / interpret predictions made by the model (“black box”).
* Tend to perform best with large training datasets.

**Recurrent neural networks (RNNs)** map a sequence of inputs to predicted output(s).

* Most common format is “**many-to-one**”, that maps an input sequence to one output value.
* Input at each time step sequentially updates the RNN cell’s “**hidden state**” (“**memory**”).
* After processing the input sequence, the hidden state information is used to predict the output.

RNNs often struggle to process long input sequences. It is mathematically difficult for RNNs to capture long-term dependencies over many time steps, which is a problem for Time Series, [as sequences are often hundreds of steps](https://www.coursera.org/learn/time-series-survival-analysis/supplement/T37OF/summary-review). Another type of Neural Networks, Long short-term memory networks (LSTMs) can mitigate these issues with a better memory system

**Long short-term memory networks** share RNNs’ conceptual structure.

* LSTM cells have the same role as RNN cells in sequential processing of the input sequence.
* LSTM cells are internally more complex, with gating mechanisms and two states: a hidden state and a cell state.

Long short-term memory networks regulate information flow and memory storage.

* LSTM cells share forget, input, and output gates that control how memory states are updated and information is passed forward.
* At each time step, the input and current states determine the gate computations.

**LSTMs vs RNNs**

LSTMs are better suited for handling long-term dependencies than RNNs. However, they are much more complex, requiring many more trainable weights. As a result, LSTMs tend to take longer to train (slower backpropagation) and can be more prone to overfitting.

These are some guidelines on how to choose LSTMs or RNNs in a Forecasting task:

Always consider the problem at hand:

* If sequences are many time steps long, an RNN may perform poorly.
* If training time is an issue, using a LSTM may be too cumbersome.
* Graphics processing units (GPUs) speed up all neural network training, but are especially recommended when training LSTMs on large datasets.

**Survival Analysis**

**Survival Analysis** focuses on estimating the length of time until an event occurs. It is called ‘survival analysis’ because it was largely developed by medical researchers interested in estimating the expected lifetime of different cohorts. Today, these methods are applied to many types of events in the business domain.

Examples:

* How long will a customer remains on books before churning
* How long until equipment needs repairs

**Survival Analysis** is useful when we want to measure the risk of events occurring and our data are Censored.

* This can be referred to as failure time, event time, or survival time.
* If our data are complete and unbiased, standard regression methods may work.
* Survival Analysis allows us to consider cases with incomplete or censored data.

The **Survival Function** is defined as S(t)=P(T>t)*S*(*t*)=*P*(*T*>*t*) . Itmeasures the probability that a subject will survive past time *t*.

This function:

* Is decreasing (non-increasing) over time.
* Starts at 1 for all observations when *t=0*
* Ends at 0 for a high-enough *t*

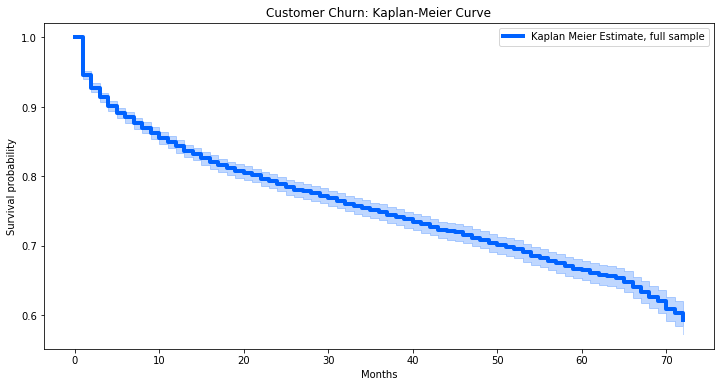
The **Hazard Rate** is defined as:

h(t)=\frac{f(t)}{S(t)}*h*(*t*)=*S*(*t*)*f*(*t*)​

It represents the instantaneous rate at which events occur, given that it has not occurred already.

The cumulative hazard rate (sum of h(t)*h*(*t*) from *t = 0* to *t = t*) represents accumulated risk over time.

**The Kaplan-Meier** estimator is a non-parametric estimator. It allows us to use observed data to estimate the survival distribution. The Kaplan-Meier Curve plots the cumulative probability of survival beyond each given time period.



Using the Kaplan-Meier Curve allows us to visually inspect differences in survival rates by category. We can use Kaplan-Meier Curves to examine whether there appear to be differences based on this feature.

To see whether survival rates differ based on number of services, we estimate **Kaplan-Meier** curves for different groups.

**Survival Analysis Approaches**

The **Kaplan-Meier approach** provides sample averages. However, we may want to make use of individual-level data to predict survival rates.

Some well-known Survival models for estimating Hazard Rates include these **Survival Regression** approaches. These methods:

* Allow us to generate estimates of total risk as a function of time
* Make use of censored and uncensored observations to predict hazard rates
* Allow us to estimate feature effects

Although these methods use time, these methods are not generally predicting a time to an event, rather predicting survival risk (or hazard risk) as a function of time.

–       The **Cox Proportional Hazard (CPH) model**

This is one of the most common survival models. It assumes features have a **constant proportional impact** on the hazard rate.

For a single non-time-varying feature *X*, the hazard rate h(t)*h*(*t*) is modeled as:

h(t)=\beta\_0(t)e^{\beta\_1X}*h*(*t*)=*β*0​(*t*)*eβ*1​*X*

\beta\_0(t)*β*0​(*t*) is the time-varying ***baseline hazard****,* and e^{\beta\_1X}*eβ*1​*X* is the (constant) proportional adjustment to the baseline hazard due to ***X***.

Using the CPH model, we can plot estimated survival curves for various categories.

–       **Accelerated Failure Time (AFT) models** (several variants including the **Weibull AFT model**)

These models differ with respect to assumptions they make about the hazard rate function, and the impact of features.