# 生存分析

——以某电信公司的电信客户流失数据为例

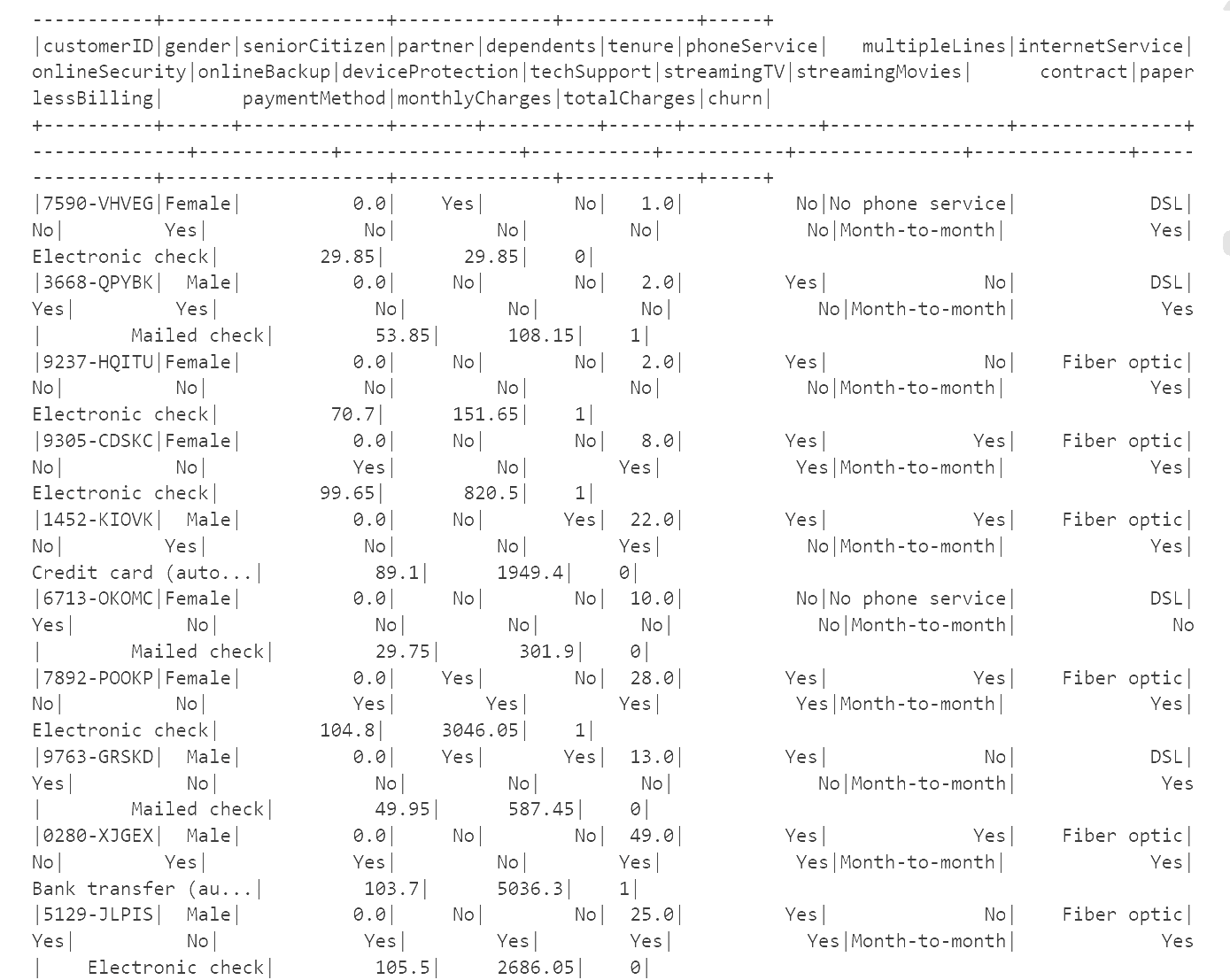
生存分析是一组统计方法，用于研究和预测直到某个感兴趣事件发生所需的时间。这种分析形式起源于医疗保健领域，最初主要关注死亡时间，但是后来可以用于各种不同的领域，如电信行业应用，以下为三个例子：

1.客户留存：人们普遍认为留存客户的成本低于获取新客户的成本。将服务取消作为感兴趣的事件，电信公司可以通过生存分析更有效地管理客户流失问题，以便更好地预测特定客户在何时可能面临流失风险。

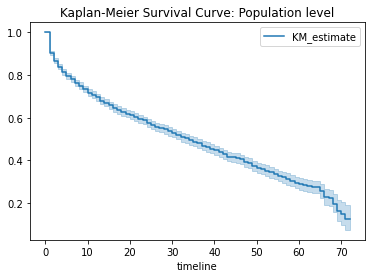
2.硬件故障：客户对产品和服务的体验质量在其决定是否续约或取消服务时起着关键作用。网络本身是这种体验的核心。将硬件故障时间作为感兴趣的事件，生存分析可用于预测硬件何时需要维修或更换。

3.设备和数据套餐升级：在客户生命周期中有一些关键节点，客户会对套餐进行更改。将套餐更改作为感兴趣的事件，生存分析可用于预测此类更改何时会发生，然后采取行动对客户选择的产品或服务产生积极影响。

面对数据，我们首先进行数据预处理，原始数据为青铜表，处理后的数据为白银表，我们将数据中churnString 列转换为数值类型的 churn 列，Yes 转换为 1，No 转换为 0，其他情况标记为 Unknown，方便我们筛选出是否进行服务的客户来进行分析，大致数据如下（使用spark进行处理）。

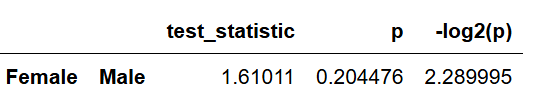
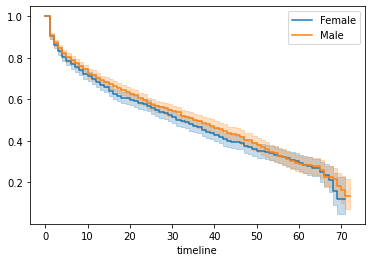


然后我们开始用卡普兰 - 迈耶法来进行构建生存概率曲线，卡普兰 - 迈耶法是一种用于构建生存概率曲线的统计方法。该方法考虑了删失数据，因而克服了使用均值或中位数时会出现的低估生存概率的问题。使用 Lifelines 进行卡普兰 - 迈耶分析时，第一步是拟合模型。这一步需要两个参数：客户在网时长（tenure）和是否流失（churn）。客户在网时长（Tenure）是指客户与公司保持业务关系的时长（如果仍为用户），或者在流失之前与公司保持业务关系的时长，而是否流失（Churn）这数据是一个布尔值，表明客户当前是否仍为用户。我们可以看一下拟合后的曲线，见下图：

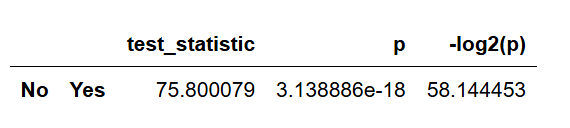
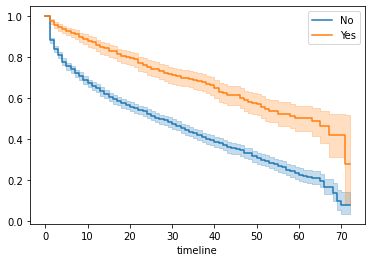


图中可以看到，客户至少存活 0 个月的概率是 100%。这由下方图表中的点 (0, 1.0) 表示。沿着生存曲线向下到中位数（34 个月），可以说，假设客户已经存活了 33 个月，那么他们至少存活 34 个月的概率为 50%。生存概率曲线周围的浅蓝色边框代表置信区间。区间越宽，置信度越低。如下方图表所示，随着时间线的增加，对估计值的置信度会降低。

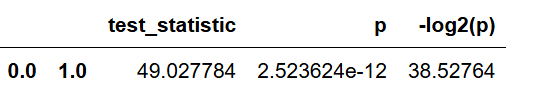
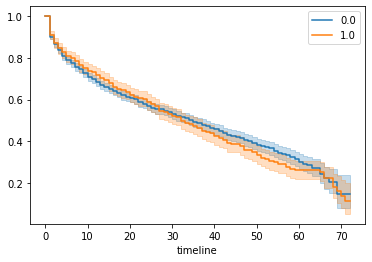
接着在协变量层面查看卡普兰 - 迈耶曲线，我们希望两条线有直观地区别，这样方便我们去研究影响客户存活的因素（接下来是对不同因素的分析）：



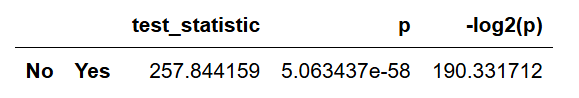
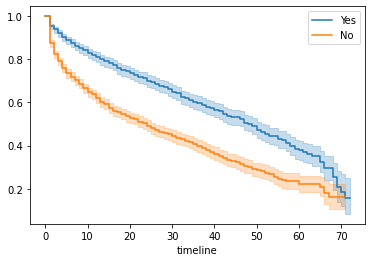
上图看出性别就没有什么差异，而且p 值大于 0.05，因此，我们不能拒绝这两个群组在统计上等效的原假设。



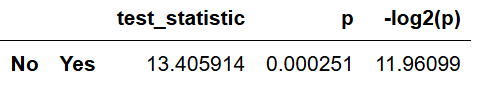
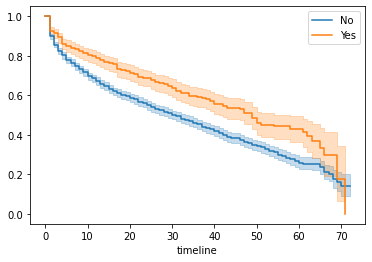
上图是有关是否有在线安全的，有在线安全的生存概率明显大于无在线安全，p值小于0.05，能拒绝原假设



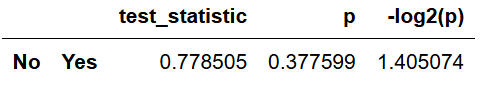
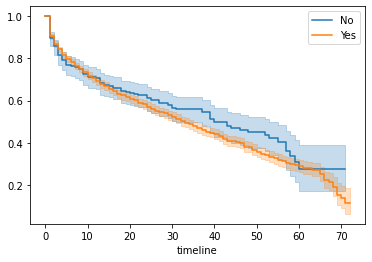
上图有关老年客户，0为不是，1为是，可以看出几乎重合，说明无差异，p值小于0.05，能拒绝原假设。



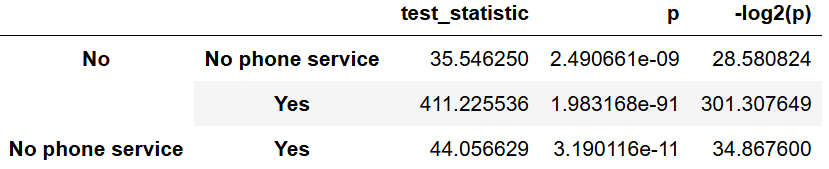
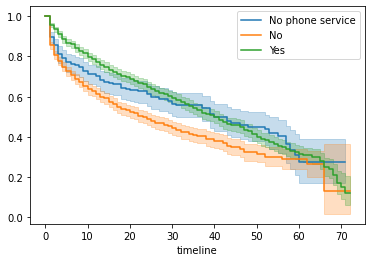
上图有关是否有伙伴，有伙伴明显留存概率比无伙伴大，p值小于0.05，能拒绝原假设



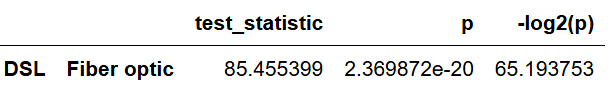
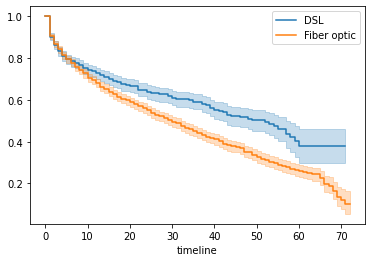
上图有关依赖性，有依赖性的留存概率比无依赖性的略大，p值小于0.05，能拒绝原假设。



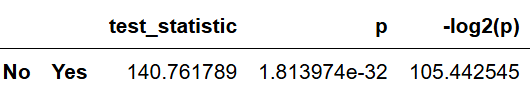
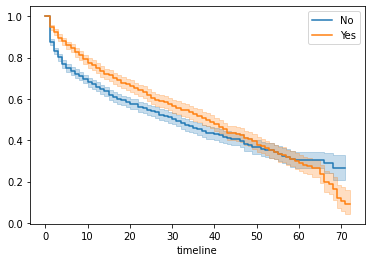
上图有关是否有电话服务发现没有电话服务的留存概率比有电话服务的略高，p值大于0.05，不能拒绝原假设。



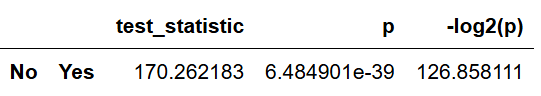
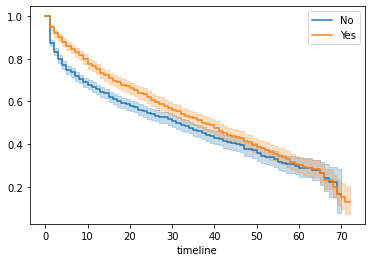
上图为多条线，为有电话服务，无电话服务，没有打电话，发现重合率较高，但p值小于0.05，能拒绝原假设



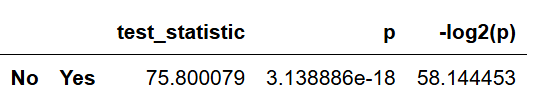
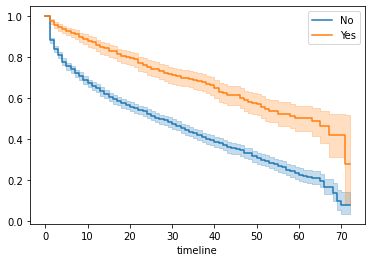
上图为网络服务方式，DSL明显高于Fiber optic，p值小于0.05，能拒绝原假设



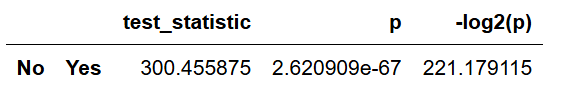
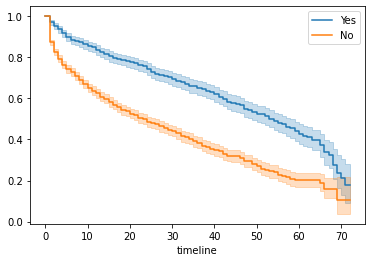
上图为是否有电视流，重合率较高，p值小于0.05，能拒绝原假设



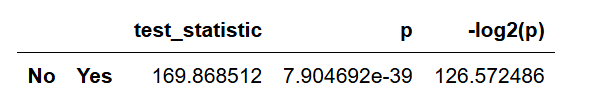
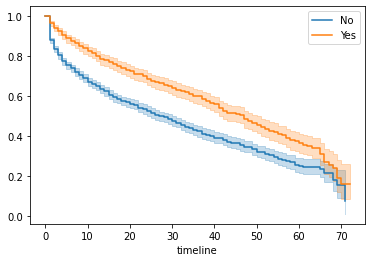
上图为是否有电影留，重合率较高，p值小于0.05，能拒绝原假设



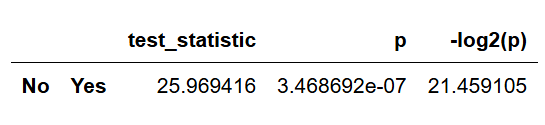
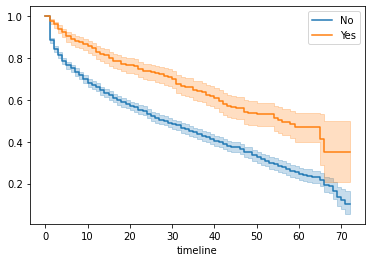
上图为是否有在线服务，明显有在线服务留存概率高，p值小于0.05，能拒绝原假设。



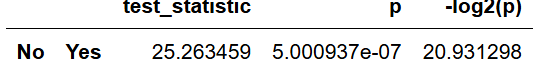
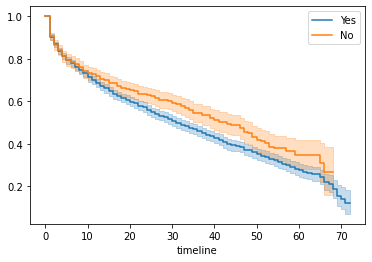
上图为是否有在线备份，有在线备份的留存概率明显比没有在线备份高，p值小于0.05，能拒绝原假设。



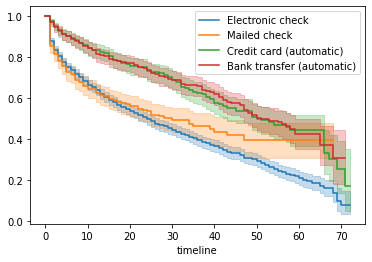
上图为是否有设备保护，有设备保护的明显比没有设备保护的留存概率高，p值小于0.05，能拒绝原假设。



上图为是否有技术支持，有技术支持的明显比没有技术支持留存概率高，p值小于0.05，能拒绝原假设。

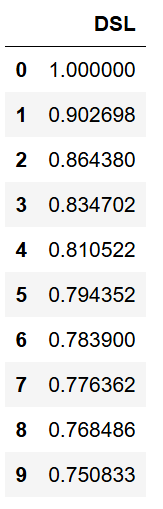


上图为是否是无纸化账单，留存概率没有太大区别，可能有纸账单留存概率更高一些，p值小于0.05，能拒绝原假设



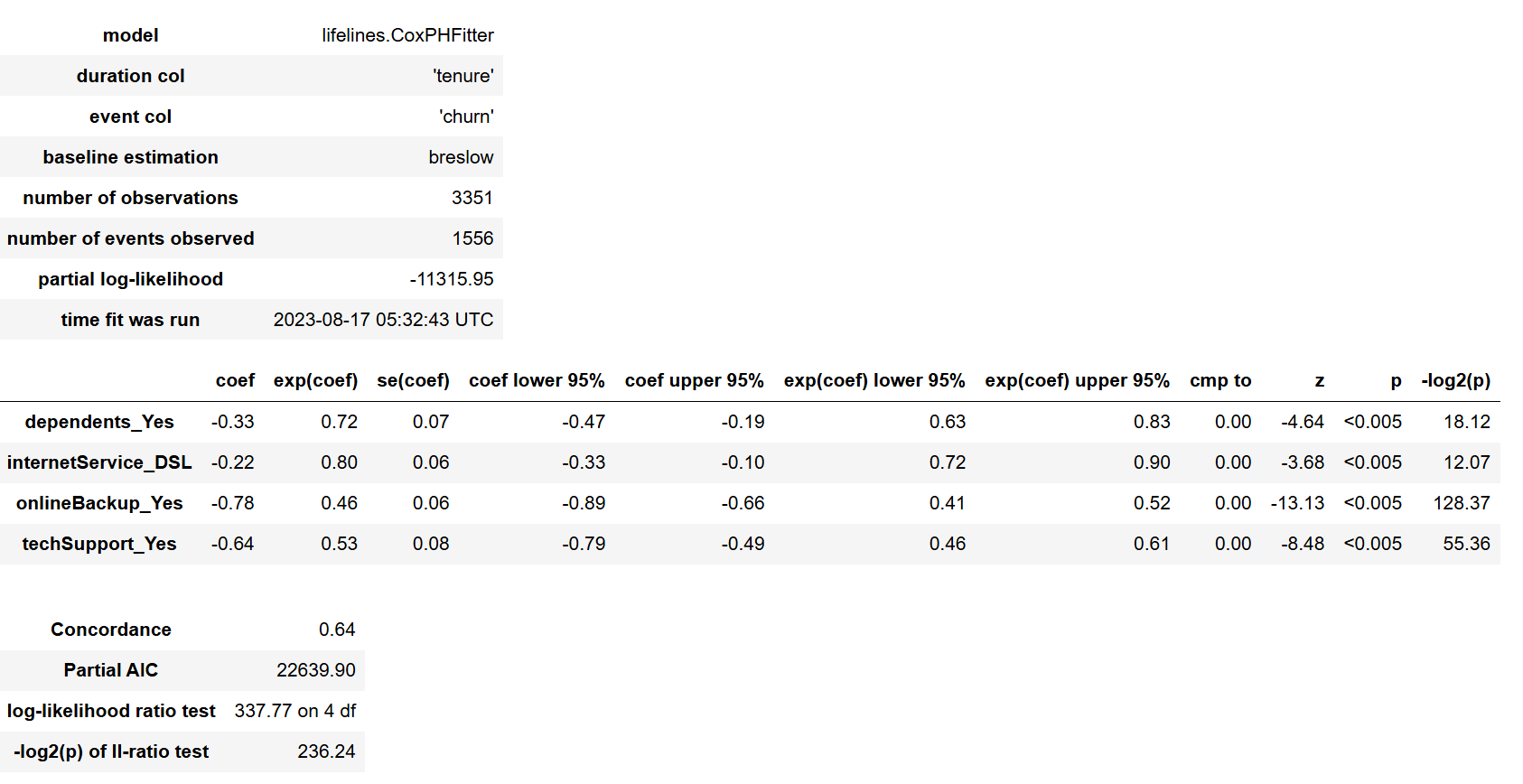
最后为支付方式，用银行账户或者卡的留存概率更高。

此外我们还可以对某一个变量进行查看他的留存概率，例如对于internetService = DSL。

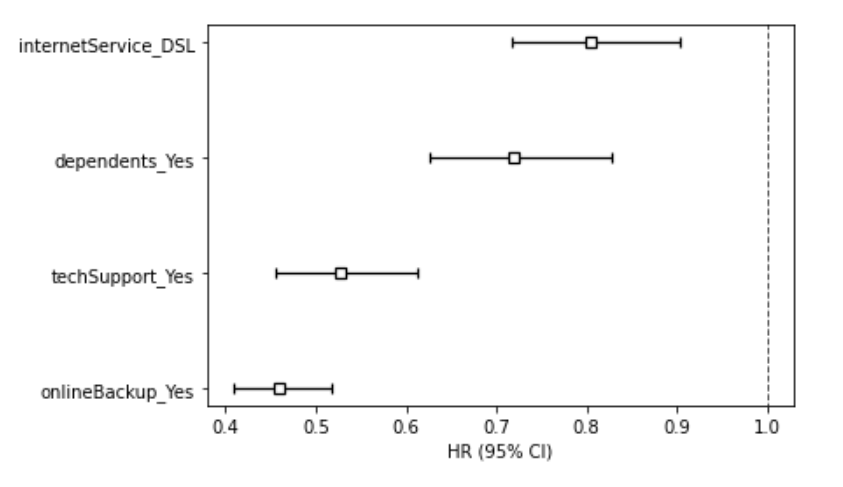
可以看到9月前留存概率75%以上。

卡普兰 - 迈耶法在单变量分析中极为有用，因为它有助于对现有数据建立更直观的理解。然而在多变量分析，我们需要用到另一个方法，即Cox风险比例模型，它是采用逆向思维的方式，通过评估风险，反推留存概率。Cox 比例风险方程表明，风险比是两项的乘积：基线风险和部分风险。

我们来看一下该模型拟合结果（数据预处理同上），见下图：

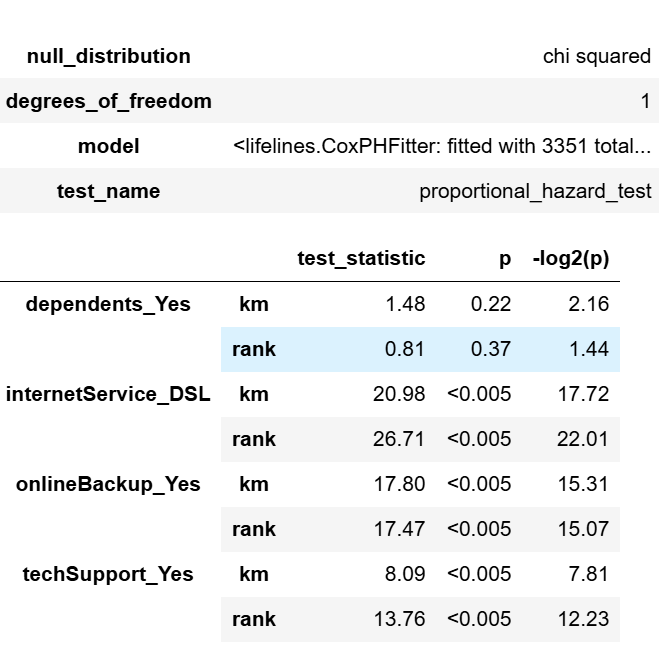


在下面标记为 “p” 的列中，可以看到每列的 p 值都小于 0.005。因此，每一列在统计上都是显著的，可以放心纳入分析。以 internetService\_DSL 为例，如下所示，系数（coef）= -0.22，指数化系数（exp (coef)） = 0.80。回顾 Cox 比例风险方程，这意味着当客户选择 DSL 作为互联网服务时，其风险比（与基线相比）降低了 0.80 倍。以下还有箱线图：



我们可以看到每个系数和指数化系数（exp (coefficient)）的上下限。

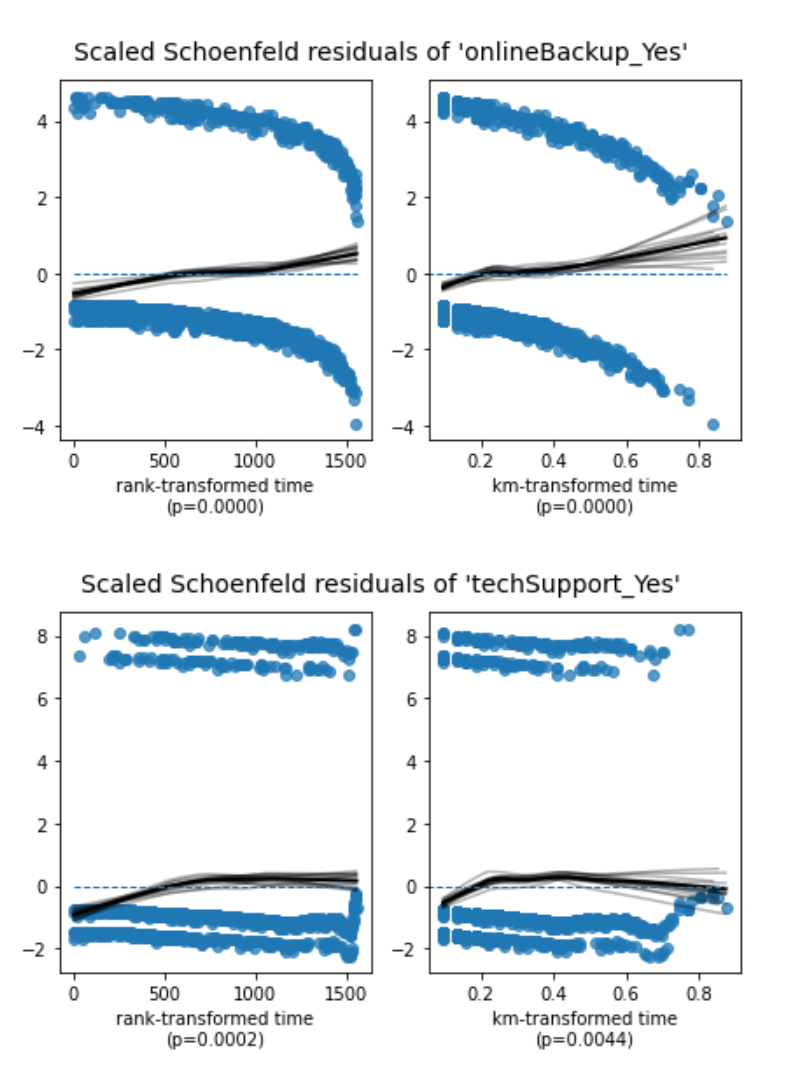
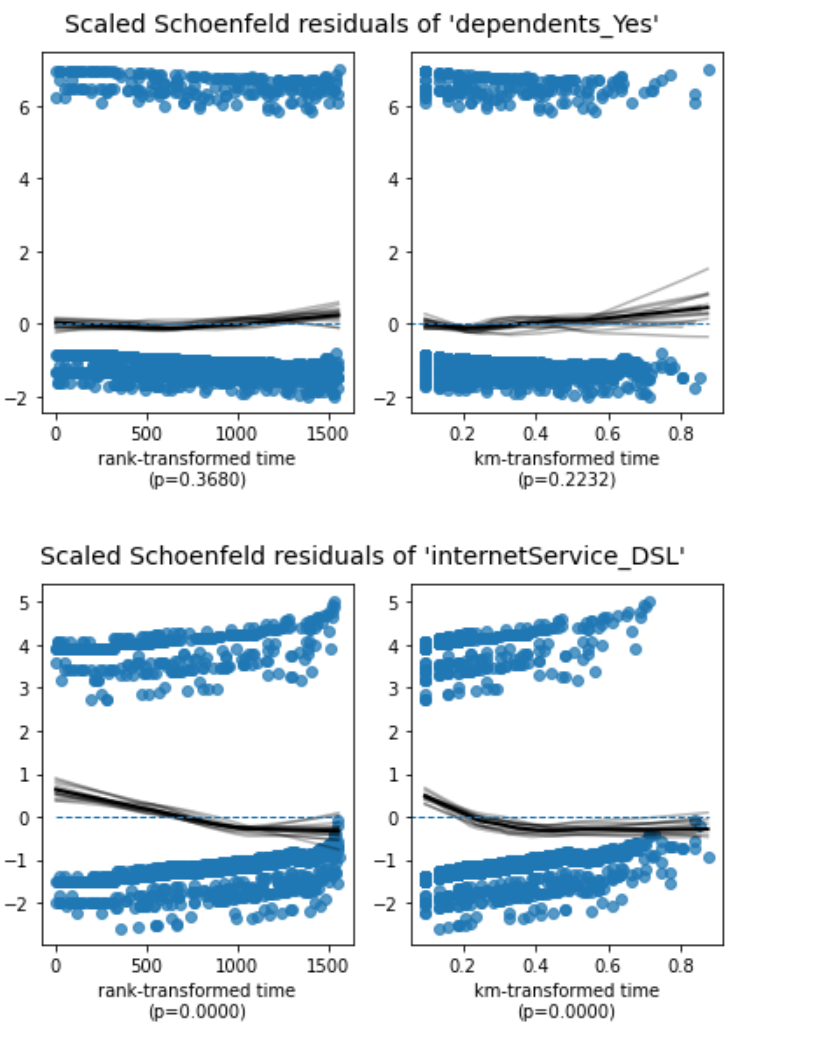
在看完我们模型结果后我们要评估他是否符合比例风险假设，因此我们采用三种方法：统计检验，舍恩菲尔德残差，对数卡普兰迈耶图。



Python建议如下：变量 “internetService\_DSL” 未通过非比例性检验：p 值小于 5e - 05。由于该变量唯一值很少（只有 2 个），你可以在.fit调用中加入strata=['internetService\_DSL', ...] 。详见下方链接 [E] 中的文档说明。

变量 “onlineBackup\_Yes” 未通过非比例性检验：p 值小于 5e - 05。由于该变量唯一值很少（只有 2 个），你可以在.fit调用中加入strata=['onlineBackup\_Yes', ...] 。详见下方链接 [E] 中的文档说明。变量 “techSupport\_Yes” 未通过非比例性检验：p 值为 0.0002。由于该变量唯一值很少（只有 2 个），你可以在.fit调用中加入strata=['techSupport\_Yes', ...] 。详见下方链接 [E] 中的文档说明。

接着我们可以看一下舍恩菲尔德残差：



在上面的输出中，每个变量都有两个图。这两个图的区别在于残差值的显示顺序：秩变换时间和卡普兰 - 迈耶变换时间。对于我们的模型，这两种图之间没有明显差异。

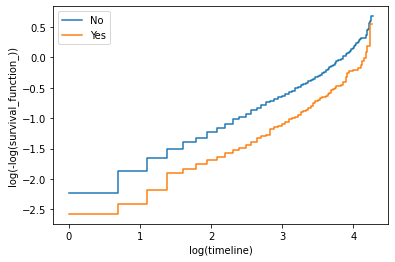
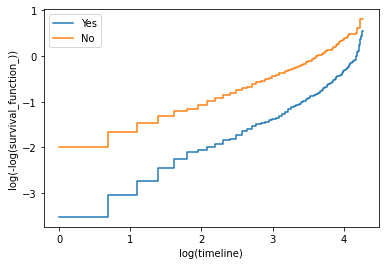
解读这些图的方式与解读线性回归的残差图类似。换句话说，在查看这类图时，我们不希望在残差中看到任何类型的模式。当不存在模式时，中间的黑线将相对平坦，表明残差与时间不相关。

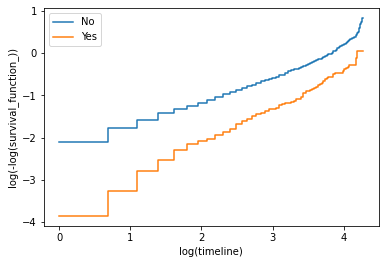
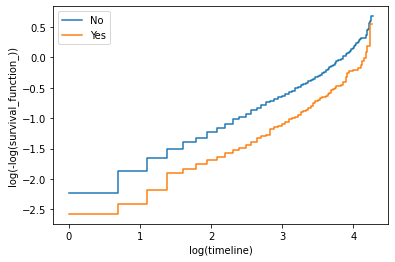
internetService\_DSL：随时间有明显且一致的趋势。

onlineBackup\_Yes：在这三个变量中趋势最为明显。

techSupport\_Yes：随时间能看到一些模式，但最显著的影响来自时间线末尾附近的点。

最后是对数卡普兰迈耶图检验：



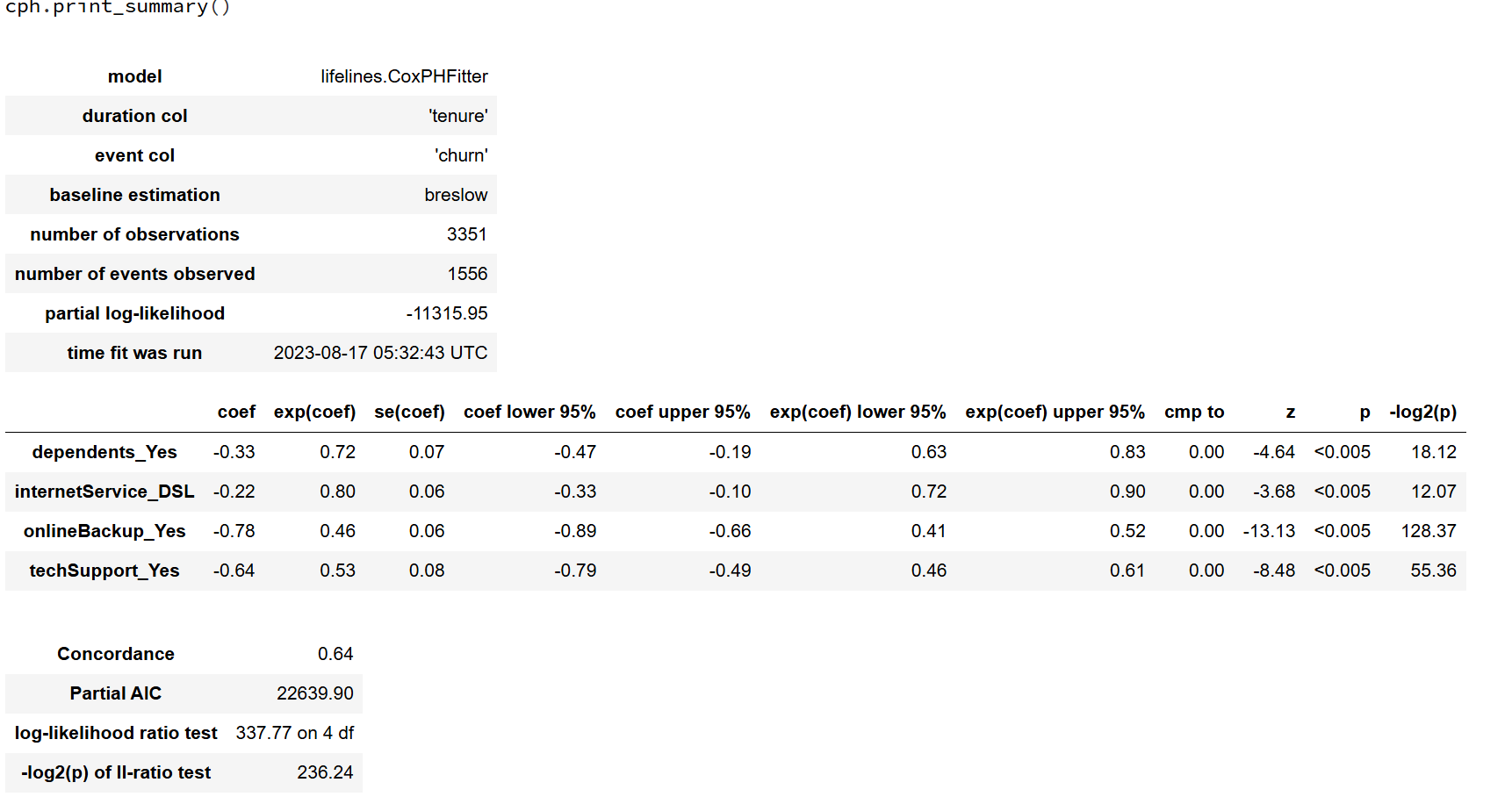


当比例风险假设未被违反时，对数 - 对数图中的卡普兰 - 迈耶曲线将呈现平行状态。这与上面 “比例风险假设” 部分所查看的图是一致的。

除了 “互联网服务”（internetService）这个变量外，从下面的图中可以看出，当对数时间轴（log (timeline)）在 1 到 3 之间时，卡普兰 - 迈耶曲线大多是平行的，但当对数时间轴小于 1 或大于 3 时，平行度就没那么高了。

接下来我们用到的是一个全新的模型，加速失效时间模型（灵感来源于人类的寿命是狗的七倍）。与卡普兰 - 迈耶法（Kaplan - Meier）和 Cox 比例风险模型不同，加速失效时间模型是一种参数模型。这意味着假设结果变量遵循特定的分布。参数模型通常比非参数和半参数模型的 “灵活性” 差，但当你能够确定结果变量的分布时，它可能是一个不错的选择。

我们依旧按照之前的数据预处理来进行拟合模型，以下是模型信息：

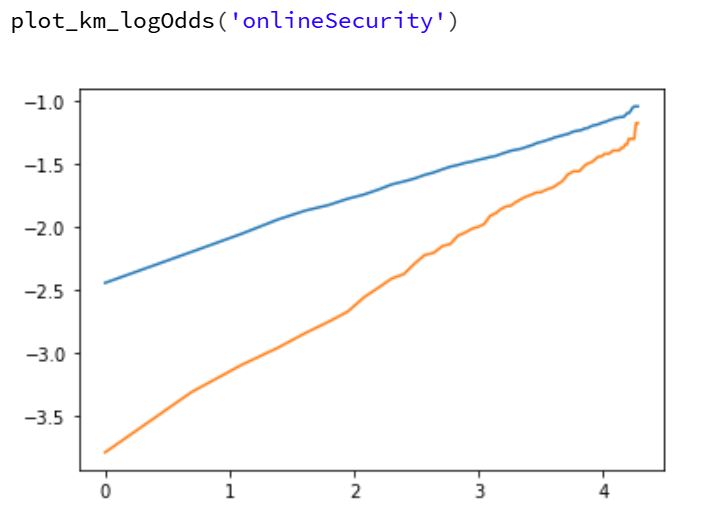
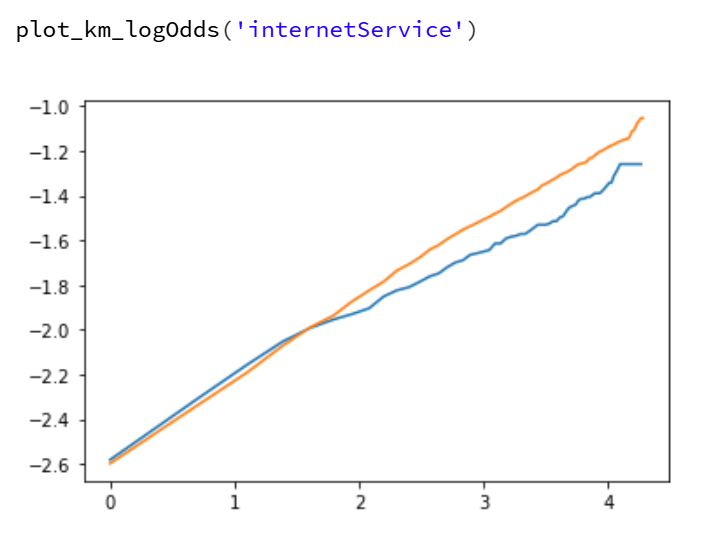
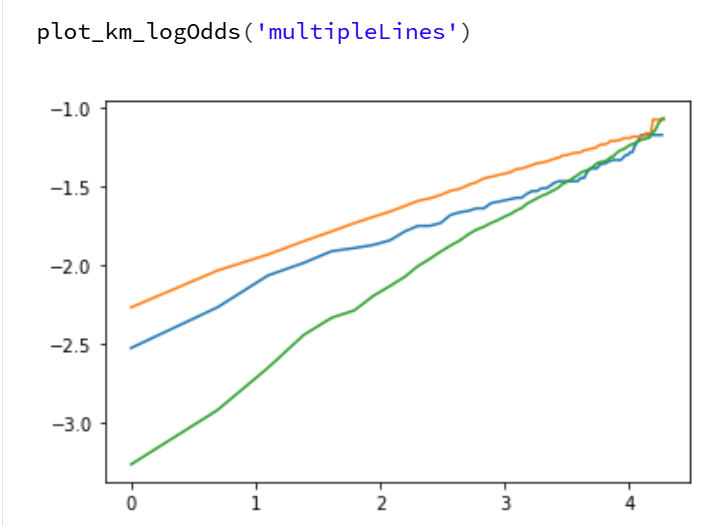
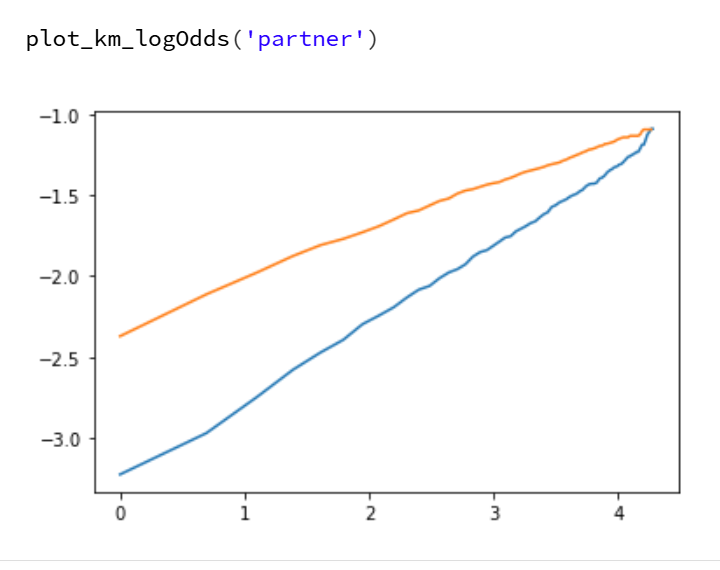


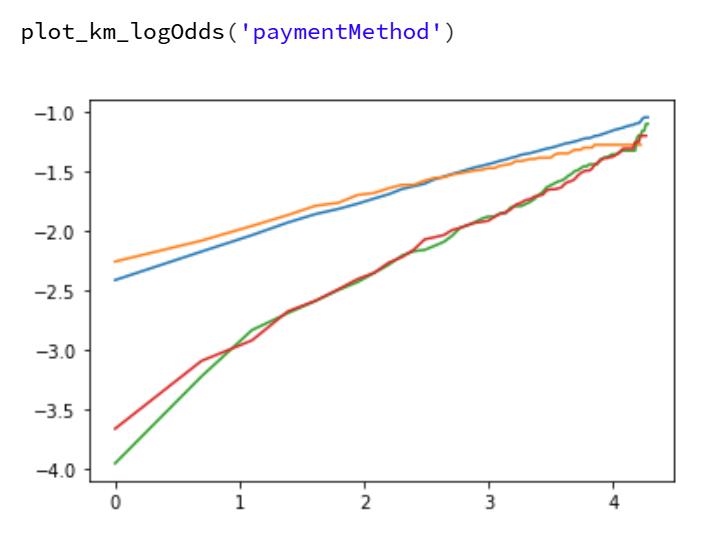
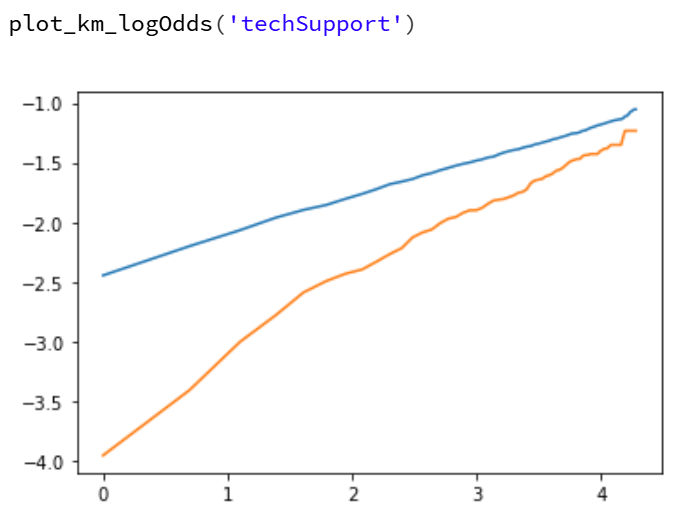
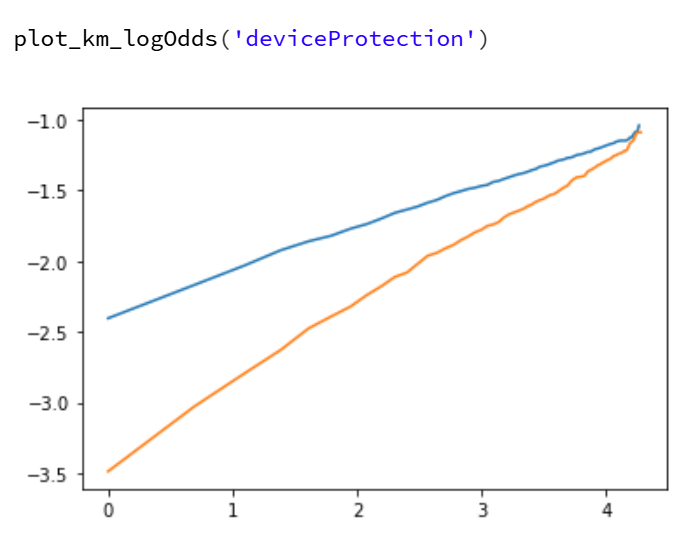
在下面标记为 “p” 的列中，可以看到每列的 p 值都小于 0.005。因此，每一列在统计上都是显著的，可以放心纳入分析。以 internetService\_DSL 为例，如下所示，系数（coef） = 0.38，指数化系数（exp (coef)） = 1.47。回顾加速失效时间方程，这意味着当客户选择光纤作为互联网服务时，其到流失的时间会加速 1.47 倍。

我们在 Cox 比例风险模型中所看到的，对数 - 对数图对于评估模型是否违反假设非常有帮助。对于 Cox 比例风险模型，对数 - 对数图是以对数时间（log (time)）为 x 轴，生存函数的对数 - 对数为 y 轴绘制的。对于加速失效时间（Accelerated Failure Time）模型，y 轴所使用的公式取决于为结果变量指定的分布类型。由于我们这里使用的是对数 - 逻辑分布，y 轴的公式为：log (1 - 生存函数) / 生存函数。

在使用加速失效时间模型时，有两个基本假设需要评估：模型是否符合比例优势假设？—— 当图中的线条平行时，答案为是。指定的分布对此模型是否合适？—— 当线条为直线时，答案为是。

以下是一些结果：





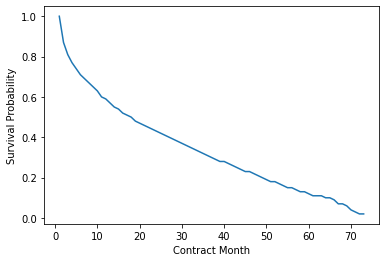
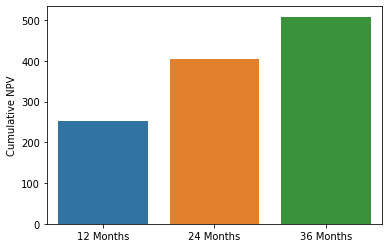
可以看到，在大多数情况下，每个图中的线条相对较直。虽然存在一些偏差，但总体情况不算糟糕。这意味着选择对数 - 逻辑分布作为指定分布是一个合理的选择。但是在大多数情况下，每个图中的线条并不平行。这意味着加速失效时间模型不适用于指定的模型。

最后为了更直观地分析和展示结果，创建仪表盘。先设置一系列下拉框和文本框组件，让用户可以选择不同的变量值和输入内部收益率。根据用户选择的值，计算并展示生存概率、每月预期利润、预期利润的净现值等指标。



上图便是一个简单的例子，根据用户选择的值，计算并展示生存概率、每月预期利润、预期利润的净现值等指标。

然后再图表展示，帮助企业确定获取客户的最大成本；生存概率曲线图表展示客户在不同时间段的留存概率。



以上就是生存分析的全部步骤，我们用到了单变量分析，风险模型拟合，加速时间模型拟合，并重新设置了不同变量的输入量，来帮助企业确定该如何投入能够更好的留住用户。生存分析是一个很有实际意义的统计方法！