Session 5C People-Oriented Technology - Experience Sharing About the New Tehcnology Applications on Domestic and International Insurance Industries

Moderator:
Hanjie Sun, FSA, FCAA

Presenter:
Ziqian Huang, Ph.D.

SOA China Symposium

28 May 2018

SOCIETY OF ACTUARIES®

科技以人为本——新科技在国内外保险业的运用经验分享

黄子谦
明德咨询

SOCIETY OF ACTUARIES®

内容纲要

- 运用新的方式去获取和整合数据
- 使用新的方法去分析数据
- 使用新的技术平台进行分析
- 科技以人文本

内容纲要

- 运用新的方式去获取和整合数据
- 使用新的方法去分析数据
- 使用新的技术平台进行分析
- 科技以人文本

数据

- 数据是公司的资产之一
- 数据是传统分析和现代机器学习的血液
- 数据的新整合和应用可以创造出新的商业价值

全方位评估被保险人风险

一个自动化的寿险和健康险承保决策引擎

内容纲要

- 运用新的方式去获取和整合数据
- 使用新的方法去分析数据
- 使用新的技术平台进行分析
- 科技以人文本

新的分析方法

- 预测性分析
  - 基于历史，预测未来
- 机器学习
  - 避免统计分布假设，通过复杂机器运算，分析数据
- 深度学习
  - 人工神经网络的最新发展
  - 分析非结构化的数据，例如文字和图像

美国疾病症状诊断预测

项目需求：
某药厂研发成功一种新药，治疗胰外分泌液不足(Exocrine pancreatic insufficiency, EPI) 。

EPI目前处于诊断不足的状态。

能否通过机器学习的方法找出没有被确诊的EPI患者。

项目挑战：
在现有的商业数据库中，被确诊是否有EPI的病人数量很少。

现有的诊断测试耗时费财。

数据库中有约3.9亿条病例，290条症状标志。

解决方案：
- 使用Python中Scikit-learn软件包。
- 评估和调节以随机森林为主的26个模型。
- 通过不同的抽样方法，结合使用标识和未标志的案例数据。
- 预留测试样本，选择最佳模型。

项目结果

- 结果显示，没有被确诊的EPI患者是在传统诊断测试中被确诊患者的人数的12倍。
- 模型显示是否有胰腺和消化方面诊断代码的病例，为是否有EPI疾病的重要预测因子。
- 客户通过使用机器学习算法分析商业数据库，可以经济，快速，准确地预测出其新药的美国市场规模。
- 学习结果对医生在实际临床诊断中有指导意义，发表在相关的医学杂志上。

国内寿险公司的退保率分析

项目需求：
需要提高现有的寿险保单退保率的预测精度。

传统的精算经验学习仍然有一定的实际应用价值和预测能力，所以客户需要有可以量化的提升指标来证明机器学习的结果比传统方法好。

项目挑战：
精算领域中的机器学习和预测性分析，对于客户来说是相对比较新的技术，对于数据的要求，分析软件的选择，硬件的配置需求，结果的解释和模型的评判，缺乏了解。

解决方案：
- 同客户业务部门协商沟通，在了解他们现有的数据处理平台和硬件架构的基础上，指导客户搭建基于Spark, Python和H2O的建模平台。
- 测试了多种不同建模方法，提交项目中的建模代码和模型评估全部代码给客户，并提供详细的文档说明。

项目结果

- 最优的机器学习模型把预测精度AUC提高34%
- 在对更多数据维度的学习中，发现了一些重要的，但从未被客户使用的内部数据字段
- 展示了某些外部宏观经济变量对预测的有用性
- 客户可以自己开始应用机器学习技术到类似的分析中

内容纲要

- 运用新的方式去获取和整合数据
- 使用新的方法去分析数据
- 使用新的技术平台进行分析
- 科技以人文本

新型平台的价值

- 新的分析和技术可以有效地实施
- 精算师可以快速扩展业务领域
- 咨询顾问可以在项目中更聚焦在自身优势的发挥

智能机器学习平台

基于Hadoop, Spark, R, Python, Livy的智能机器学习平台

概观
无需编程的预测性分析平台。使机器学习项目更加省时，让咨询和分析人员更专注在项目的目标设定，框架构建和价值传递。

数据管理
简便地整理任意大小的数据。可视化的界面允许用户通过鼠标的点击和托拉把数据载入Hadoop系统并调用Spark引擎处理数据。

模型建立
直接调用开源软件 R and Python的算法。

可视化结果
用户可以计算动态图和设计定制的结果报表。

高级功能
多个分析人员共同完成一个模型建立。
在Jupyter笔记本编程。

智能机器学习平台的价值

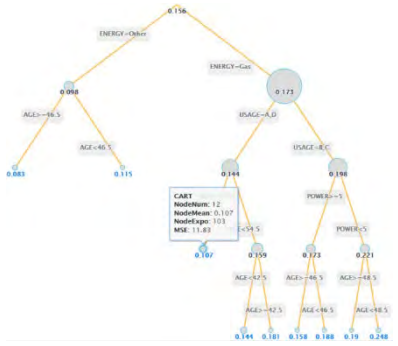
- 使用并整合高级计算机语言和开源的机器学习算法，通过可视化的用户操作界面使得编程不再是预测性分析项目实施的障碍。
- 具有建模理论基础，但缺乏编程经验的咨询顾问，在2天左右的培训后，可以给客户提供预测性分析的服务。

数据管理
节约最多 **50%**的时间

- **探索** 通过简单的分析
- **整理** 使用直观的工作流程图
- **增强** 加入外部的数据

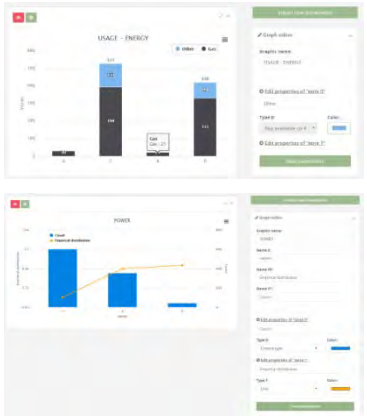
模型建立
使用可扩展性的算法

- **预测性模型** 使用开源的算法 (CART, RF, GBM, GLM, etc.)
- **可以扩展性** 基于Spark的算法



可视化
交互性的图标

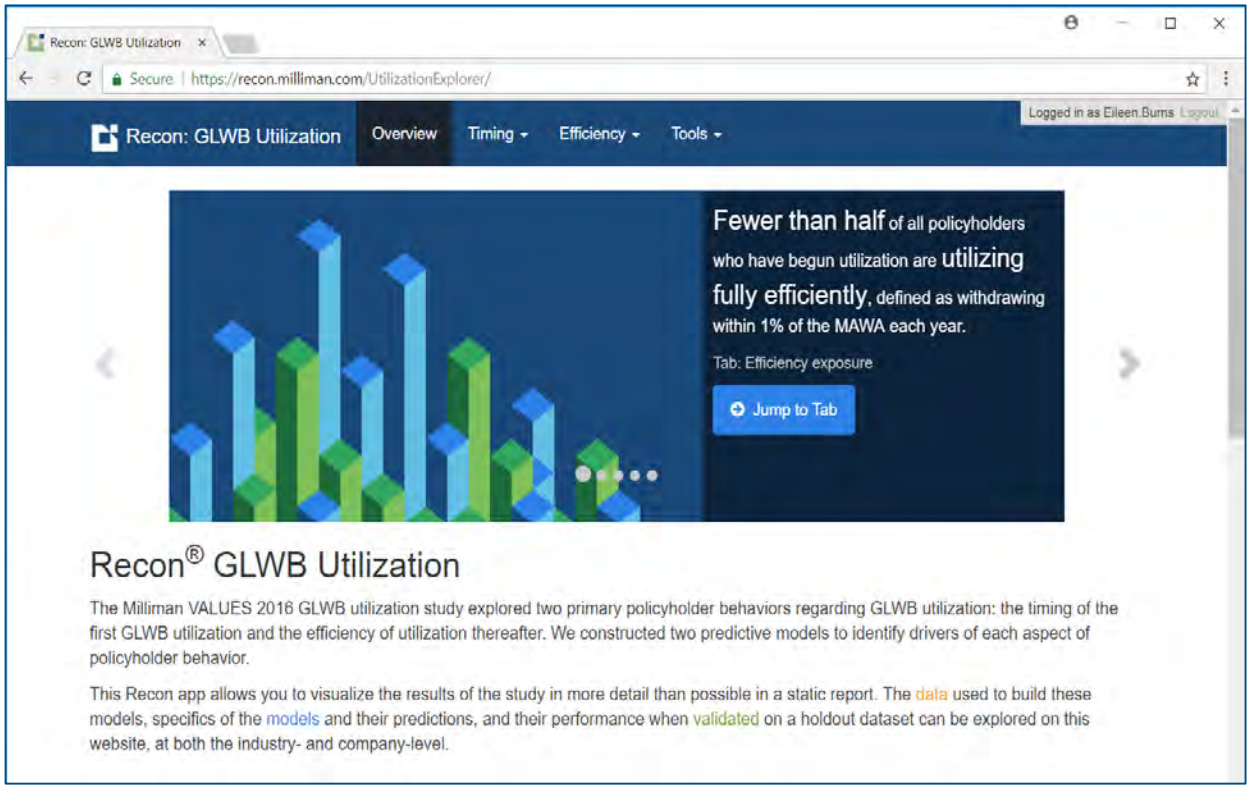
- **定制化** 结果图表导入最后报告中



软件部署
支持不同的设备

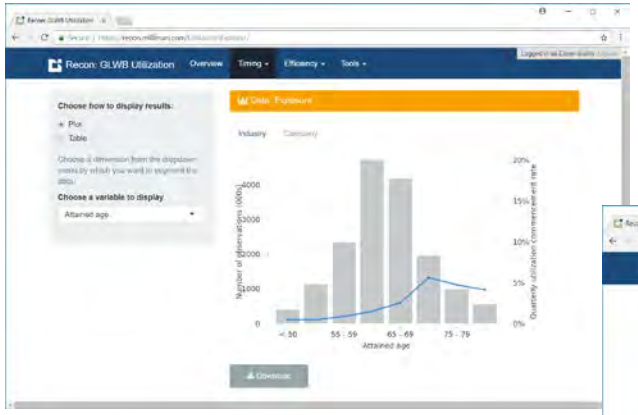
- **使用虚拟机.** 软件可以在不同的操作系统下安装
- 不同的部署模式
  - 个人电脑
  - 私有计算集群
  - 云

可视化结果展示平台

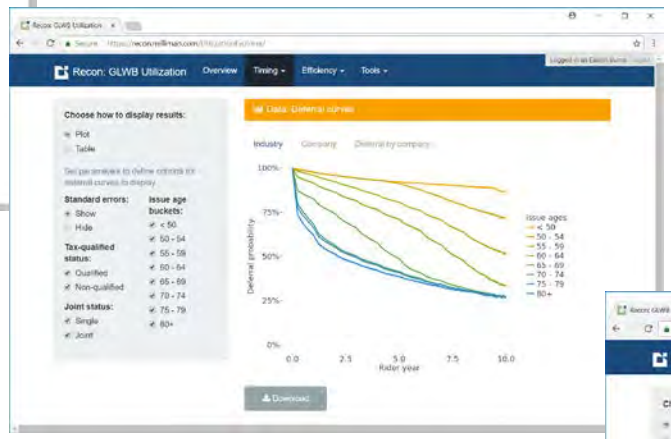


- 平台的目标是转变精算师探索有关保险客户行为经验学习的方式。
- 推动保险公司减少人为判断，采用数据驱动的决策流程。
- 列出结果亮点，让用户灵活查看结果数据。
- 新的方式来发表行业经验学习的结果，取代100多页的文本报告：更少的成词，更多的信息。

可视化结果展示平台

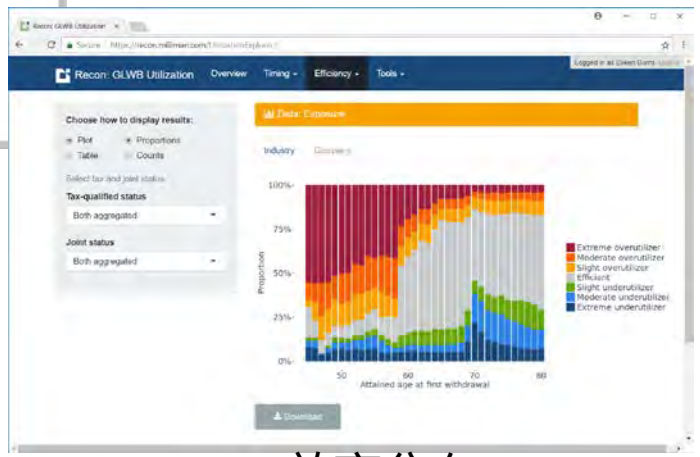


数据分布



时间分布

- 公司的结果紧邻行业的数据，方便比较。
- 用户可以根据预测模型的结果设定相应的假设。



效率分布

内容纲要

- 运用新的方式去获取和整合数据
- 使用新的方法去分析数据
- 使用新的技术平台进行分析
- 科技以人文本

新科技的应用要有可以量化的商业结果

- 越来越多的人讨论新的技术
- 不是所有的新技术都得到了成功的应用
- 新技术的应用需要前期资金投入
- 可量化的商业结果必不可少

新科技的应用要能满足或超越最终客户的需求

- 越来越多的新科技在出现和研究中
- 科技的根本目的仍然是满足或超越客户的需要
- 新科技的应用需要给最终用户带来更多的价值
  - 定价
  - 理赔
  - 反欺诈
  - 更高效的疾病诊断

谢谢!