



中国精算师协会
China Association of Actuaries



华南地区非寿险精算讲堂 第12讲

机器学习在精算实务中的应用

安永保险与精算服务总监 姚佶 博士

2017年9月

目录

- ▶ 机器学习简介  
- ▶ 机器学习在保险精算领域的应用
- ▶ 机器学习模型性能比较
- ▶ 小结与问答环节

基于大数据和机器学习的基础上，数字化人工智能将带来第四次工业革命

18世纪

第一次 工业革命

 机械

技术是通过蒸汽和水实现为第一家工厂供能

19世纪

第二次 工业革命

 电气

电力使劳动分工和大规模生产成为可能

20世纪

第三次 工业革命

 自动化

信息技术使工作编程成为可能并终结了对人力的依赖

今天

第四次 工业革命

人工智能



- 从发展脉络看，人工智能一直处于技术创新的前沿，近年来更是呈现集中爆发态势
- 在智能搜索、人工交互、可穿戴设备等领域得到了前所未有的重视
- 人工智能将成为产业界力夺的前沿领域

AlphaGo



谷歌AlphaGo机器人先后完胜韩国围棋第一人李世石和世界第一人柯洁

Google无人驾驶



穿戴式智能联网设备

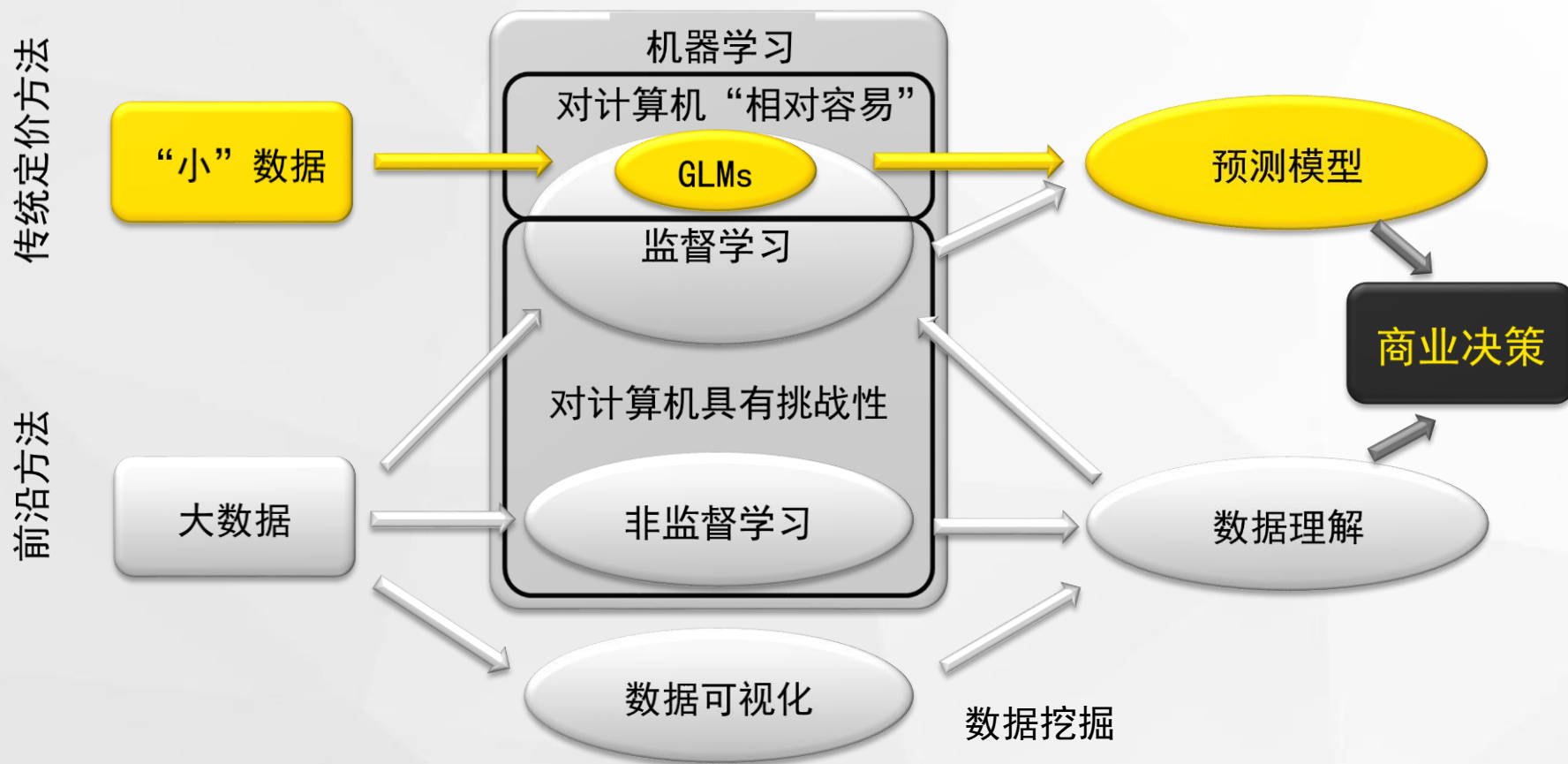


Google眼镜
运动手环
智能T恤
智能运动鞋

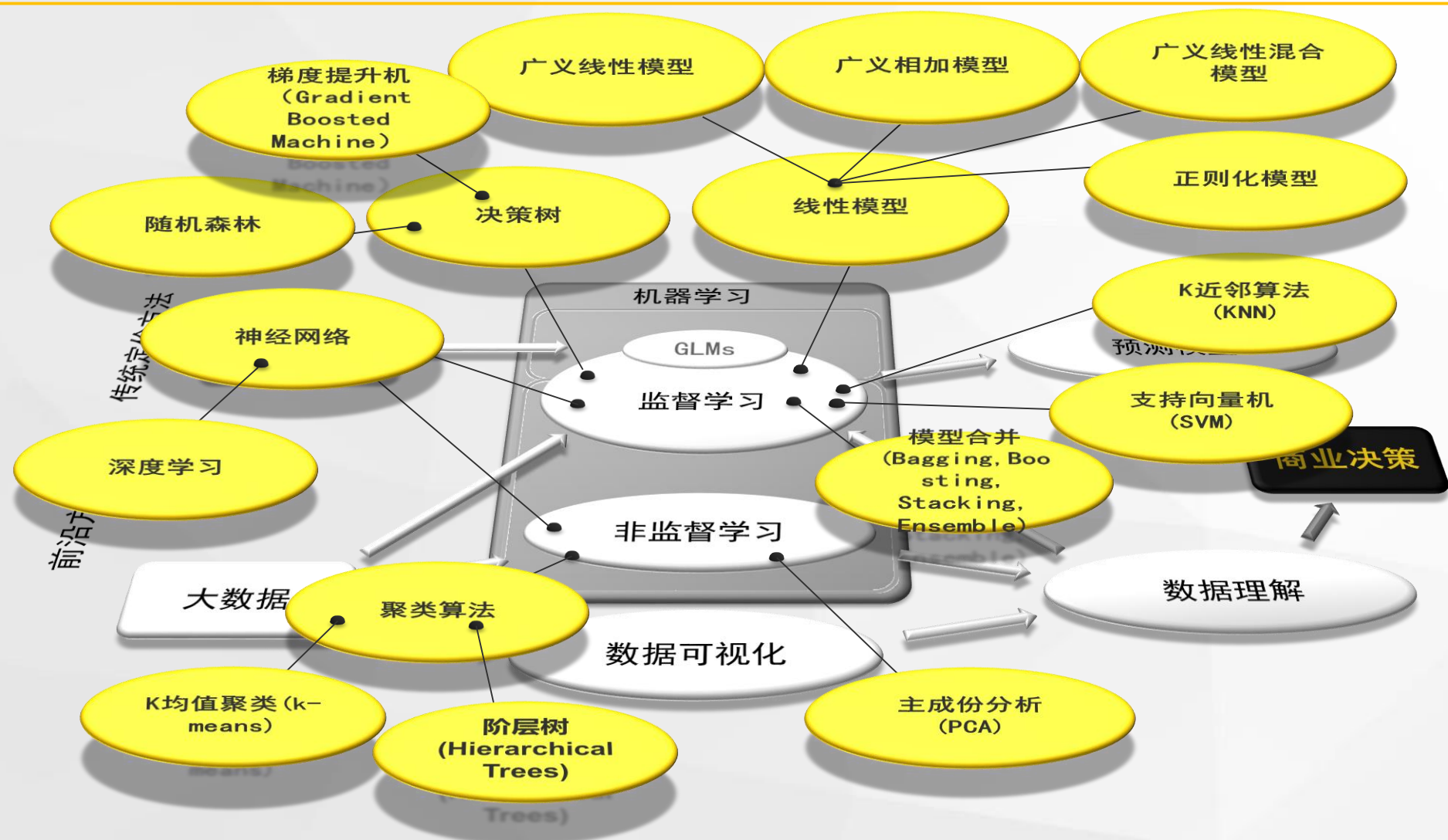
人工语音智能检索



传统定价方法适用于在数据源有限的情况下进行预测 前沿方法则更适合应用于能够获取大数据的情况下进行预测



机器学习的种类繁多，可根据不同的应用场景选择适合的算法



近年来机器学习的突破在于模型的结果大幅提升，将助推信息通信技术乃至人类社会生产生活发生深刻的革命性变化

01 由低级算法向高级算法发展

算法创新：

- ▶ 搜索巨头美国谷歌公司每天都要进行200多项改进搜索算法的在线实验
- ▶ 谷歌正在掀起一场大数据革命，陆续完成由关键字匹配到知识图谱、语义搜索的算法创新



02 由文本检索向语音图像检索发展

图像、声音识别技术提高：

- ▶ 融合了深度学习技术的搜索引擎正大幅度提升图像搜索的准确率。
- ▶ 吸纳了自然语言处理和云操作处理技术的搜索引擎，可将语音指令转化为实时搜索结果。例如Apple Siri



03 由互联网搜索向云物搜索演进

物联网技术持续发展：

- ▶ 基于人工智能的搜索引擎技术正向物联网、信息化不断深化应用。
- ▶ 基于人工智能的搜索引擎技术和云操作处理技术不断耦合，正在推动云计算技术发生重要变革。



04 自然语言交互已经成为互联网上成熟应用，蕴含无穷商机。

语音识别技术日趋成熟：

- ▶ 自然语言交互发展已较成熟，可与视觉操控、姿态操控和手势操控等人工智能感应技术结合应用。



开源编程语言的发展使得机器学习的算法更具可行性

Python: scikit-learn (<http://scikit-learn.org/stable/>)



Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.

Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices, Grouping experiment outcomes

Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency

Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning

Modules: grid search, cross validation, metrics. — Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.

Modules: preprocessing, feature extraction. — Examples

R: CRAN Task View: Machine Learning & Statistical Learning (<https://cran.r-project.org/web/views/MachineLearning.html>)

CRAN Task View: Machine Learning & Statistical Learning

Maintainer: Torsten Hothorn

Contact: Torsten.Hothorn at R-project.org

Version: 2016-06-24

Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this field of research is usually referred to as machine learning. The packages can be roughly structured into the following topics:

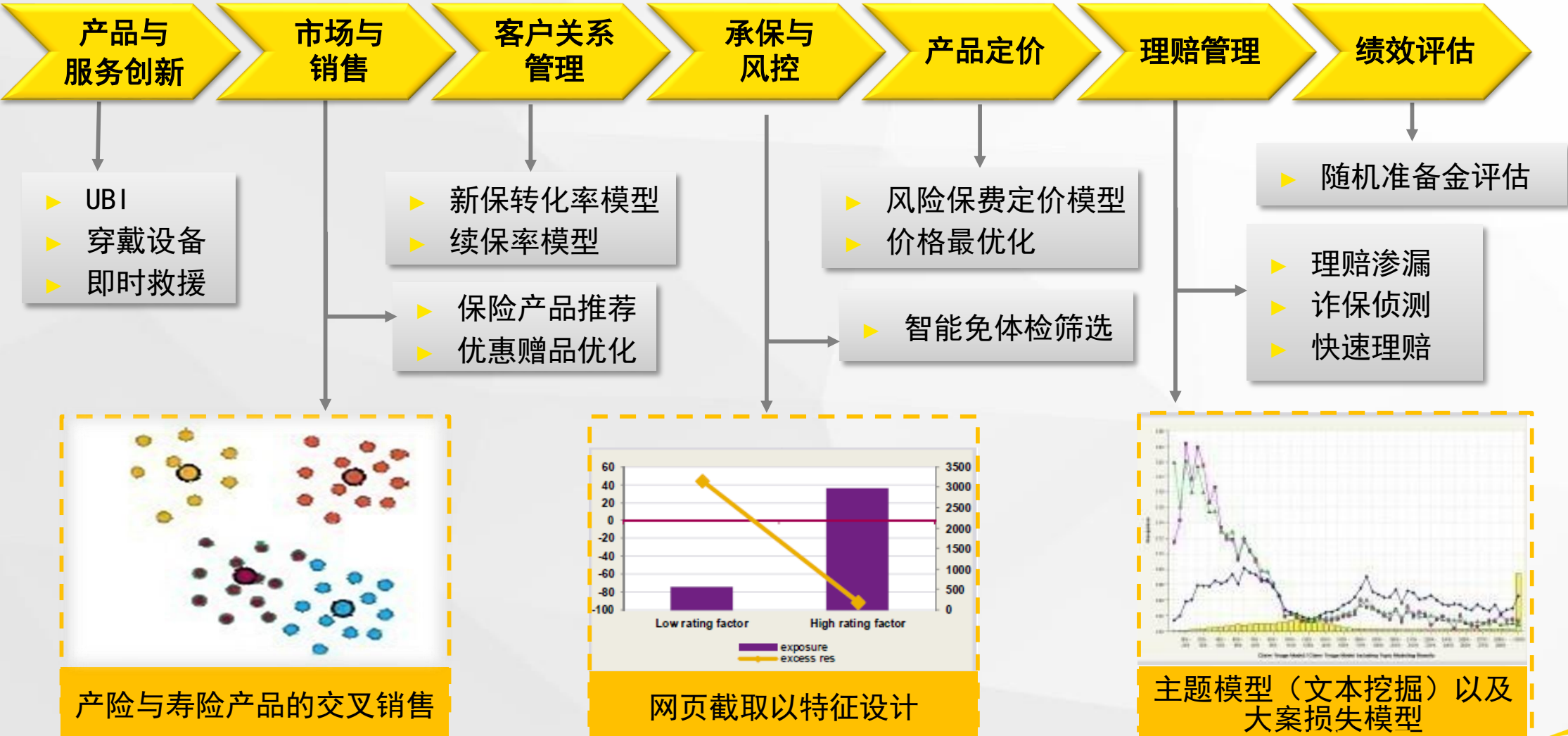
- Neural Networks** : Single-hidden-layer neural network are implemented in package [nnet](#) (shipped with base R). Package [RSNNs](#) offers an interface to the Stuttgart Neural Network Simulator (SNNS). An interface to the FCNN library allows user-extensible artificial neural networks in package [FCNN4R](#).
- Recursive Partitioning** : Tree-structured models for regression, classification and survival analysis, following the ideas in the CART book, are implemented in [rpart](#) (shipped with base R) and [tree](#). Package [rpart](#) is recommended for computing CART-like trees. A rich toolbox of partitioning algorithms is available in [Weka](#), package [RWeka](#) provides an interface to this implementation, including the J4.8-variant of C4.5 and M5. The [Cubist](#) package fits rule-based models (similar to trees) with linear regression models in the terminal leaves, instance-based corrections and boosting. The [C50](#) package can fit C5.0 classification trees, rule-based models, and boosted versions of these.
- Two recursive partitioning algorithms with unbiased variable selection and statistical stopping criterion are implemented in package [party](#). Function `ctree()` is based on non-parametrical conditional inference procedures for testing independence between response and each input variable whereas `mob()` can be used to partition parametric models. Extensible tools for visualizing binary trees and node distributions of the response are available in package [party](#) as well.
- Tree-structured varying coefficient models are implemented in package [vcpart](#).
- For problems with binary input variables the package [LogicReg](#) implements logic regression. Graphical tools for the visualization of trees are available in package [maptree](#).
- Trees for modelling longitudinal data by means of random effects is offered by package [REEMtree](#). Partitioning of mixture models is performed by [RPM](#).
- Computational infrastructure for representing trees and unified methods for prediction and visualization is implemented in [partykit](#). This infrastructure is used by package [evtree](#) to implement evolutionary learning of globally optimal trees. Oblique trees are available in package [oblique.tree](#).
- Random Forests** : The reference implementation of the random forest algorithm for regression and classification is available in package [randomForest](#). Package [ipred](#) has bagging for regression, classification and survival analysis as well as bundling, a combination of multiple models via ensemble learning. In addition, a random forest variant for response variables measured at arbitrary scales based on conditional inference trees is implemented in package [party](#). [randomForestSRC](#) implements a unified treatment of Breiman's random forests for survival, regression and classification problems. Quantile regression forests [quantregForest](#) allow to regress quantiles of a numeric response on exploratory variables via a random forest approach. For binary data, [LogicForest](#) is a forest of logic regression trees (package [LogicReg](#)). The [varSelRF](#) and [Boruta](#) packages focus on variable selection by means for random forest algorithms. In addition, packages [ranger](#) and [Rborist](#) offer R interfaces to fast C++ implementations of random forests.
- Regularized and Shrinkage Methods** : Regression models with some constraint on the parameter estimates can be fitted with the [lasso2](#) and [lars](#) packages. Lasso with simultaneous updates for groups of parameters (groupwise lasso) is available in package [grplasso](#); the [grpre](#) package implements a number of other group penalization models, such as group MCP and group SCAD. The L1 regularization path for generalized linear models and Cox models can be obtained from functions available in package [glmnet](#), the entire lasso or elastic-net regularization path (also in [elasticnet](#)) for linear regression, logistic and multinomial regression models can be obtained from package [glmnet](#). The [penalized](#) package provides an alternative implementation of lasso (L1) and ridge (L2) penalized regression models (both GLM and Cox models). Package [RXshrink](#) can be used to identify and display TRACEs for a specified shrinkage path and to determine the appropriate extent of shrinkage. Semiparametric additive hazards models under lasso penalties are offered by package [ahaz](#). A generalisation of the Lasso shrinkage technique for linear regression is called relaxed lasso and is available in package [relaxo](#). Fisher's LDA projection with an optional LASSO penalty to produce sparse solutions is implemented in package [penalized.LDA](#). The shrunken centroids classifier and utilities for gene expression analyses are implemented in package [pamr](#). An implementation of multivariate adaptive regression splines is available in package [earth](#). Variable selection through clone selection in SVMs in penalized models (SCAD or L1 penalties) is implemented in package [penalizedSVM](#). Various forms of penalized discriminant analysis are implemented in packages [hda](#), [rda](#), and [sda](#). Package [Liblinear](#) offers an interface to the LIBLINEAR library. The [ncvreg](#) package fits linear and logistic regression models under the SCAD and MCP regression penalties using a coordinate descent algorithm. High-throughput ridge regression (i.e., penalization with many predictor variables) and heteroskedastic effects models are the focus of the [bigRR](#) package. An implementation of bundle methods for regularized risk minimization is available from package [bmrn](#). The Lasso under non-Gaussian and heteroskedastic errors is estimated by [hdm](#), inference on low-dimensional components of Lasso regression and of estimated treatment effects in a high-dimensional setting are also contained. Package [SIS](#) implements sure independence screening in generalised linear and Cox models.
- Boosting** : Various forms of gradient boosting are implemented in package [gbm](#) (tree-based functional gradient descent boosting). The Hinge-loss is optimized by the boosting implementation in package [bst](#). Package

目录

- ▶ 机器学习简介
- ▶ 机器学习在保险精算领域的应用
- ▶ 机器学习模型性能比较
- ▶ 小结与问答环节



机器学习在保险精算领域具有极其广泛的应用场景，从业务前端到承保理赔端再到绩效评估，机器学习都有用武之地



海外保险公司提供数据以建模竞赛的形式解决商业问题，所采用的模型多采用机器学习的建模方式

01 Kaggle是一个数据分析的竞赛平台

企业（包括保险公司）可以将数据、问题和评价指标发布到Kaggle上，以竞赛的形式向广大的数据科学家征集解决方案。

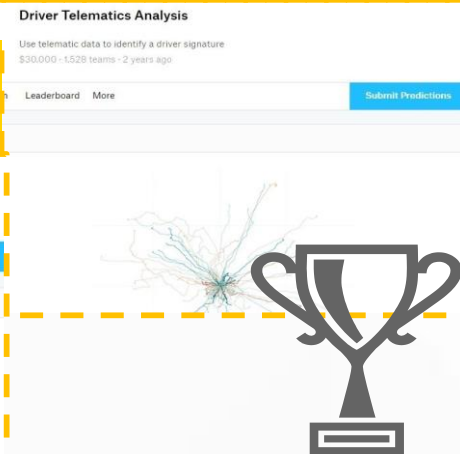
02 Kaggle上的参赛者有机会获得比赛丰厚的奖金

Kaggle上的参赛者将数据下载下来，分析数据，然后运用机器学习、数据挖掘等知识，建立算法模型，解决问题得出结果，最后将结果提交，如果提交的结果符合指标要求并且在参赛者中排名第一，会获得比赛丰厚的奖金。

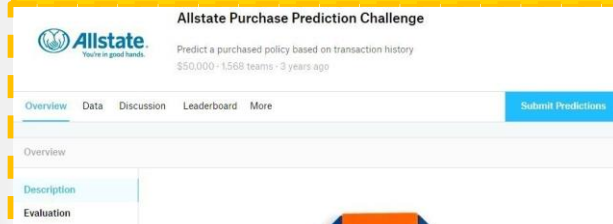


法国AXA保险公司 车载信息系统分析

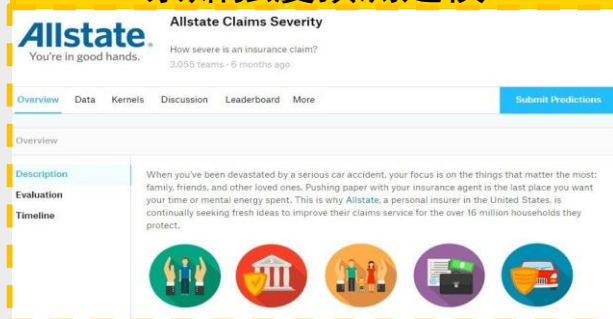
法国AXA保险公司
车载信息系统分析



美国Allstate保险公司 保单购买预测建模



美国Allstate保险公司 索赔强度预测建模



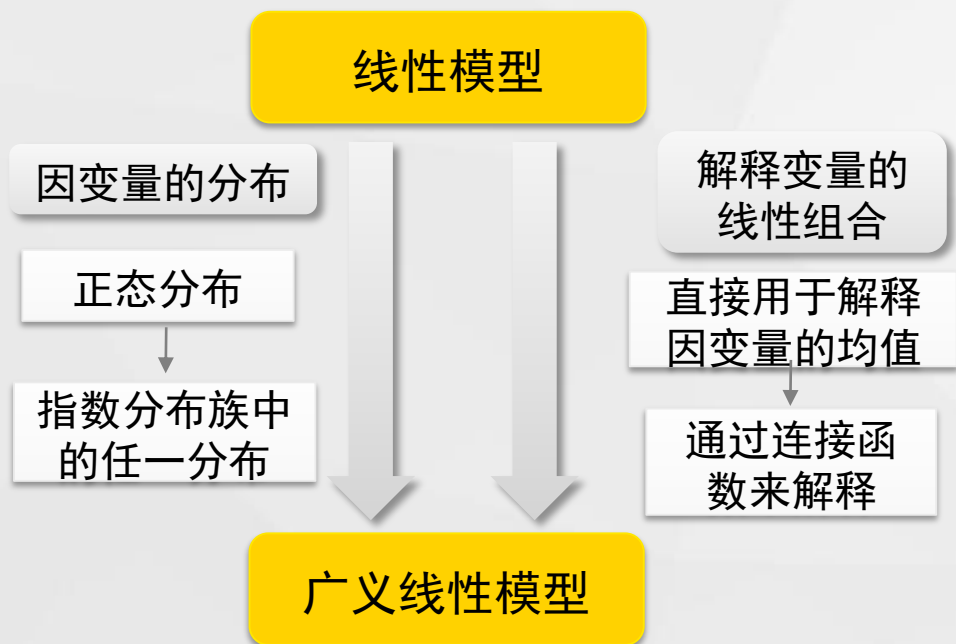
美国Liberty相互保险公司 火灾风险预测建模



我们可以从7个维度衡量机器学习方法的选择



广义线性模型（GLM）运行速度较快，稳定性较强，但预测能力一般

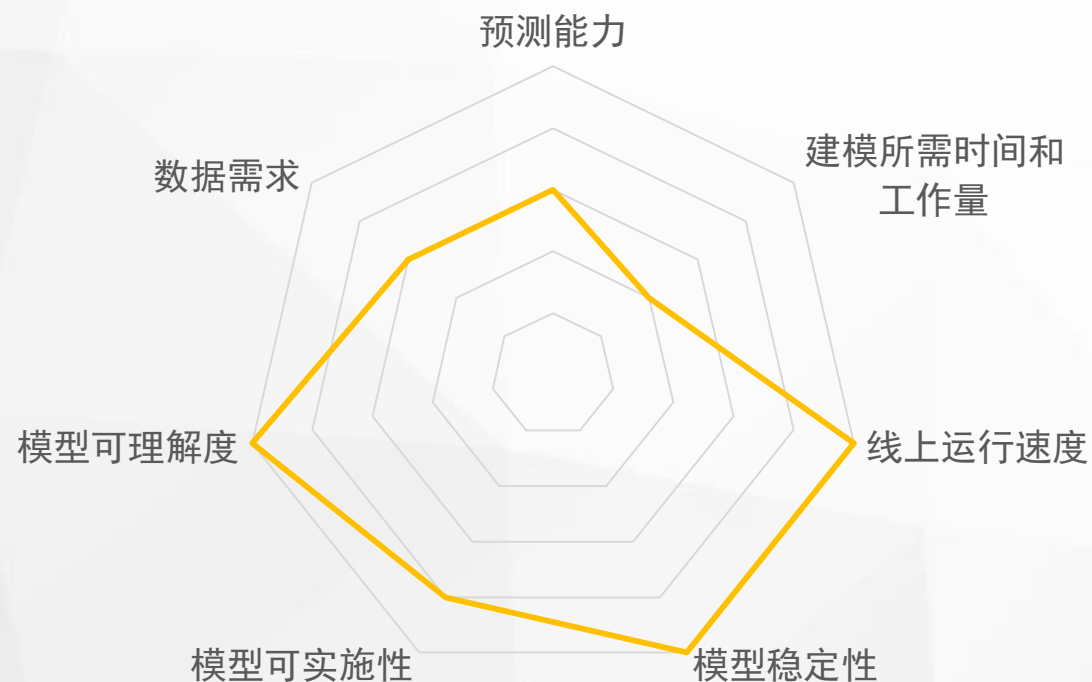


- 一般的预测函数形式为：

$$f(\underline{x}) = g^{-1}(\underline{x} \cdot \underline{\beta})$$

- 其中 β 是通过最小化损失函数 $L(\beta | X, y)$ 测算得到

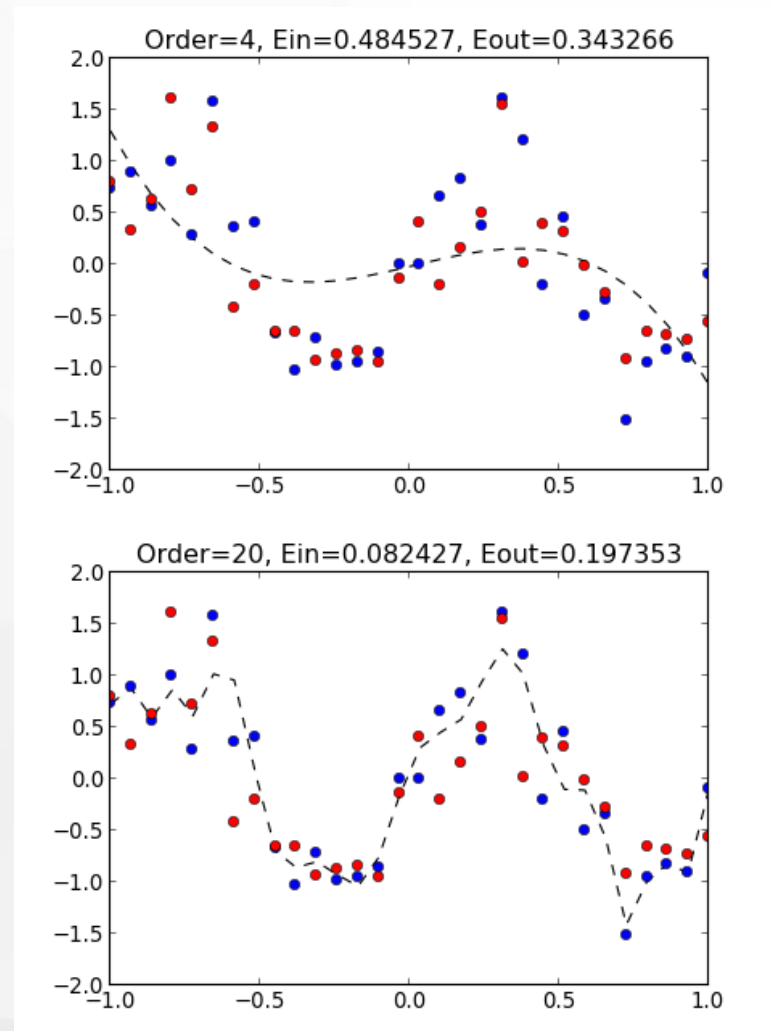
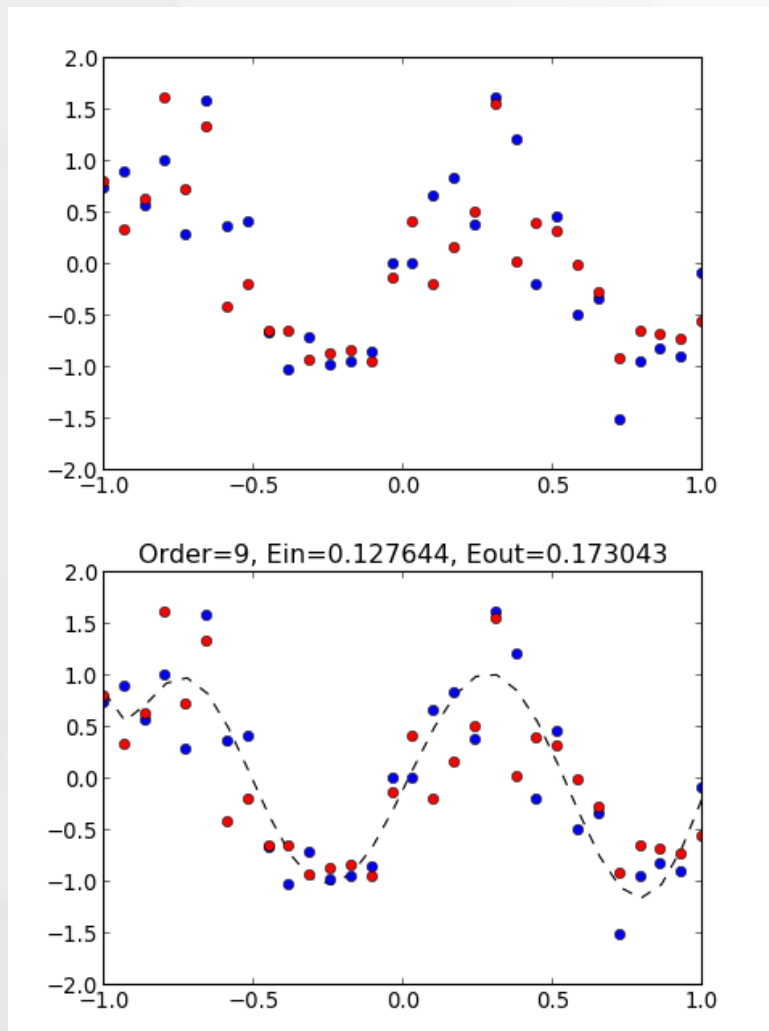
GLM模型评价



广义线性模型的过拟合问题

当模型过分拟合时，统计模型更多地预测了随机残差/噪声而不是真正的规律

图表中的蓝色点是训练数据而红色点是测试数据



惩罚回归模型运行速度较快，稳定性较强，若使用得当预测能力较广义线性模型强

广义线性模型

对似然函数加入
惩罚约束以防止
过拟合

惩罚回归模型

Ridge

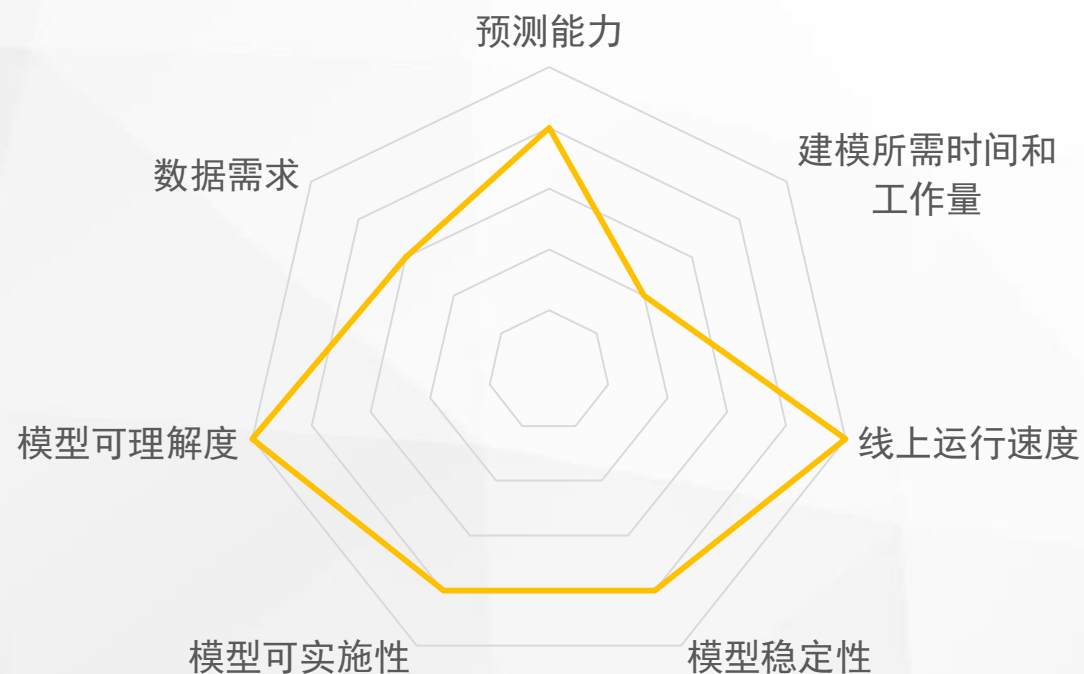
惩罚回归模型
的损失函数

$$L(\beta|X, y) + \lambda_1 \sum_i |\beta_i| + \lambda_2 \sum_i \beta_i^2$$

LASSO

弹性网络

惩罚回归模型评价



决策树模型可理解度较高，运行速度较快，但预测能力和模型稳定性较低

决策树模型呈现的是一种树形结构，可以认为是if-then规则的集合。



模型具有很好的可读性，且分类速度快



可能会产生过度匹配的问题(所以一般都会有决策树的剪枝过程)

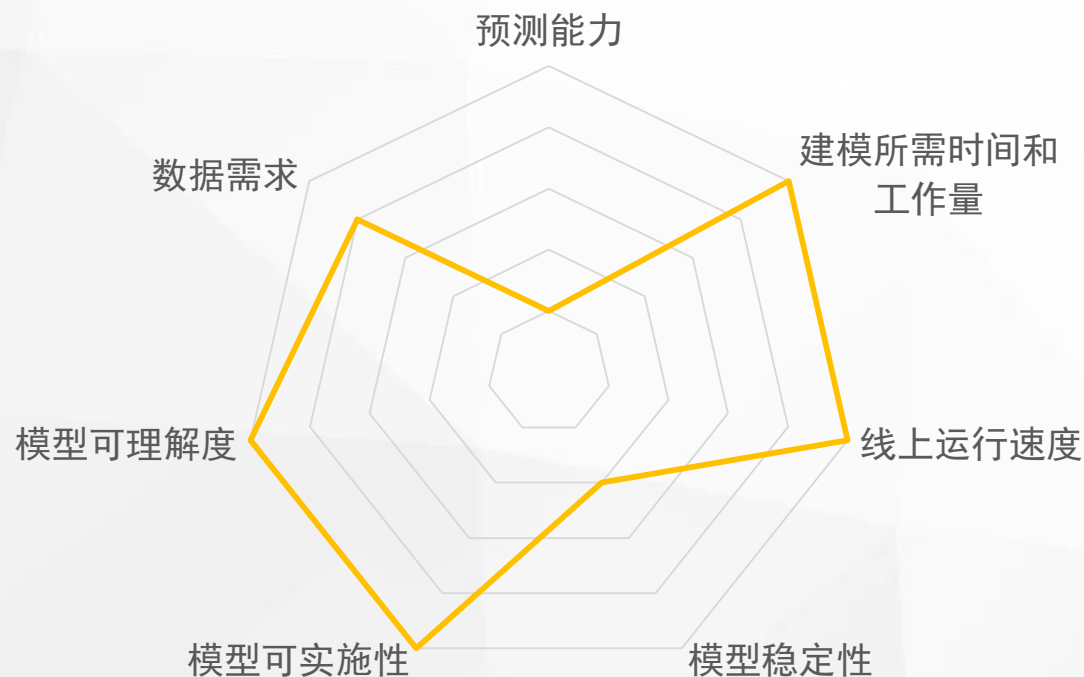
1. 特征选择

2. 决策树生成

3. 决策树修剪



决策树模型评价



梯度提升机模型（GBM）预测能力较强，但运行速度较慢，模型可理解度较低



梯度提升机模型是决策树模型的一种，该方法使子模型之间相互协作，并利用后一个子模型对前一个模型的失误进行修正

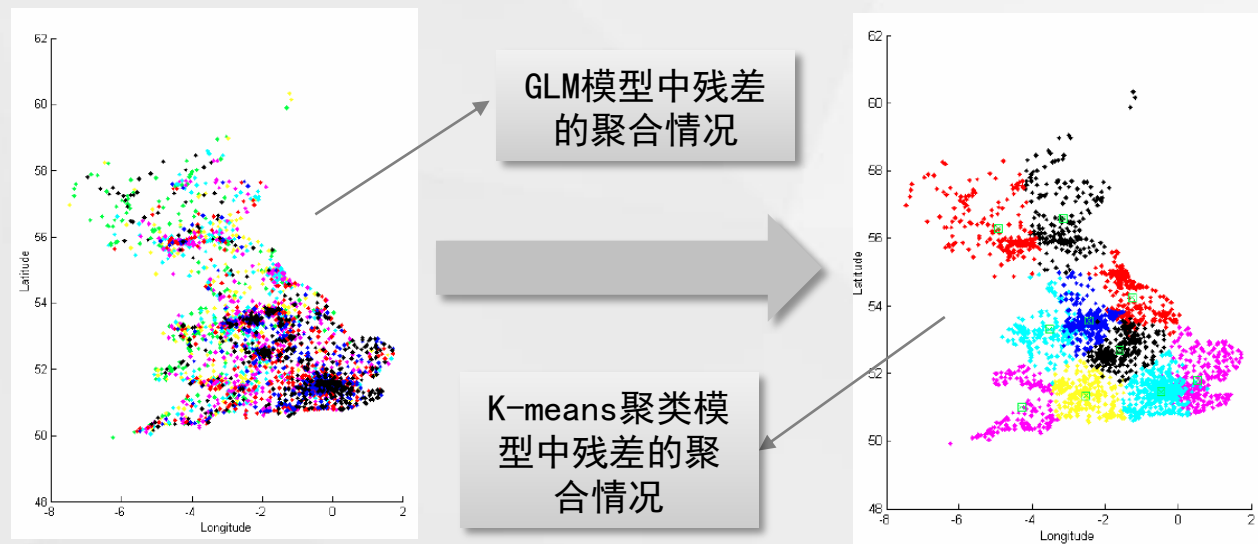
梯度提升机模型在预测的时候，对于输入的一个样本实例，首先会赋予一个初值，然后会遍历每一棵决策树，每棵树都会对预测值进行调整修正，最后得到预测的结果。

梯度提升机每一次的计算是为了减少上一次的残差(residual)，而为了消除残差可以在残差减少的梯度(Gradient)方向上建立一个新的模型。

梯度提升机（GBM）模型评价



聚类算法的预测能力较强，运行速度较快，但对数据的要求较高，模型的稳定性较低



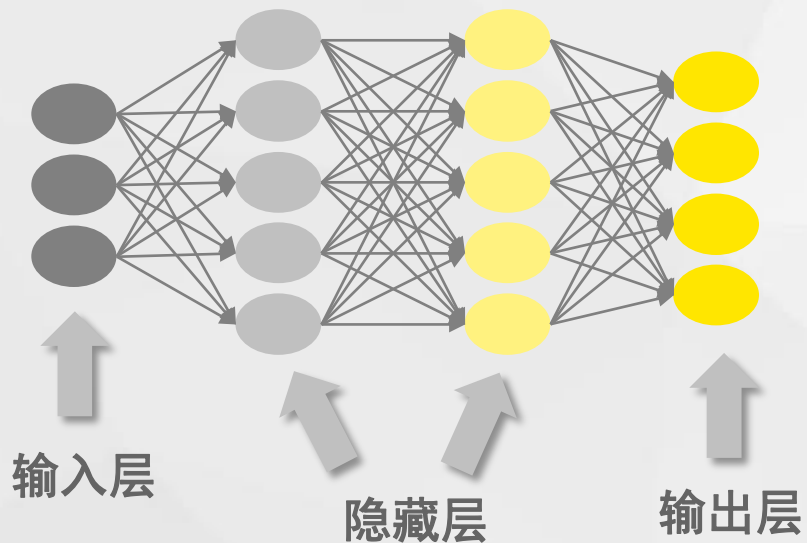
1 聚类分析又称群分析，它是研究（样品或指标）分类问题的一种统计分析方法，同时也是数据挖掘的一个重要算法。

2 聚类（Cluster）分析是由若干模式（Pattern）组成的，模式通常是一个度量（Measurement）的向量，或者是多维空间中的一个点。

3 聚类分析以相似性为基础，在一个聚类中的模式之间比不在同一聚类中的模式之间具有更多的相似性。



神经网络模型方法预测能力较强，运行速度较快，但分析所需时间和工作量较大，模型可理解度较低



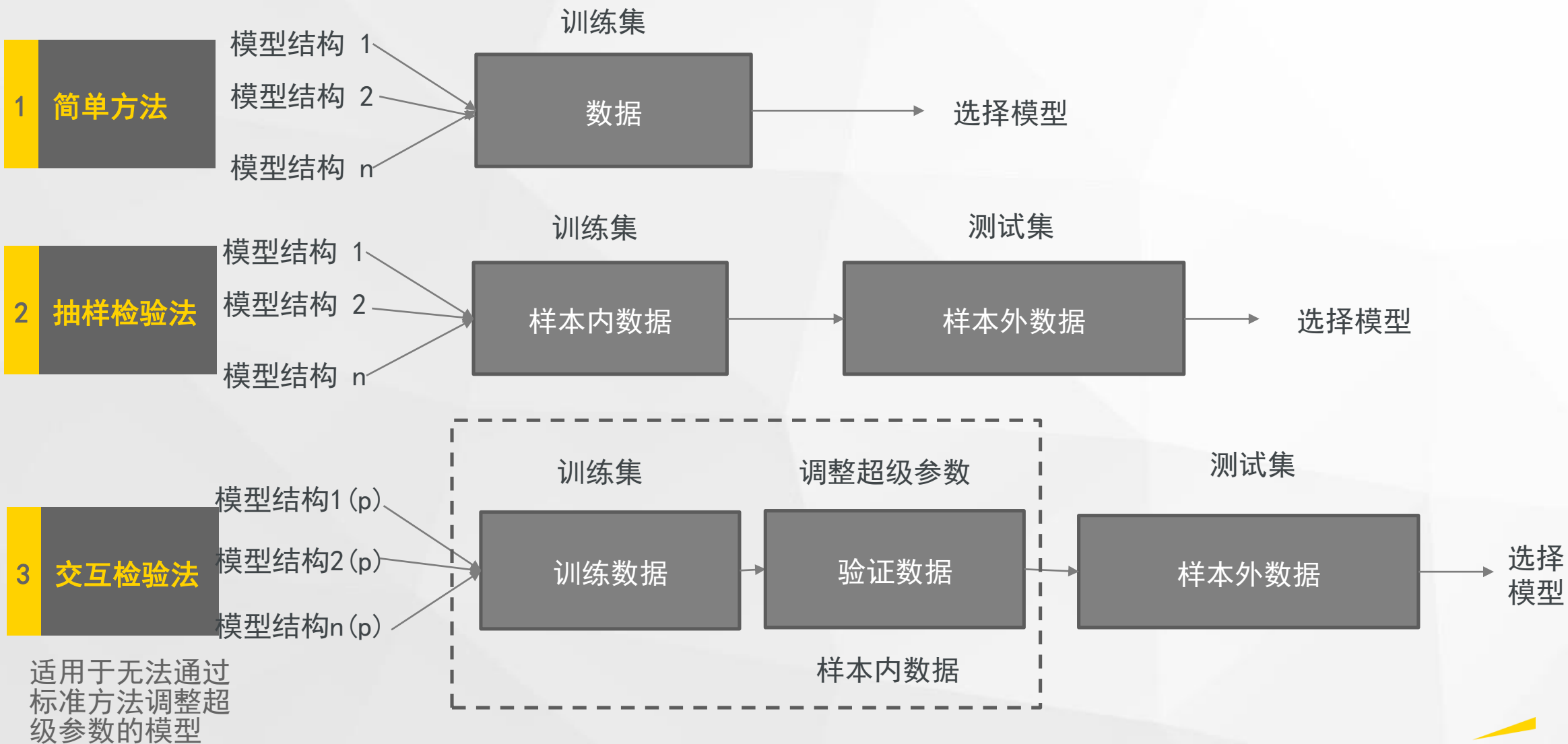
神经网络（Neural Networks, NN）是由大量的、简单的处理单元（称为神经元）广泛地互相连接而形成的复杂网络系统，它反映了人脑功能的许多基本特征，是一个高度复杂的非线性动力学习系统。

神经网络具有大规模并行、分布式存储和处理、自组织、自适应和自学能力，特别适合处理需要同时考虑许多因素和条件的、不精确和模糊的信息处理问题。

神经网络模型评价



模型中参数的选择方法

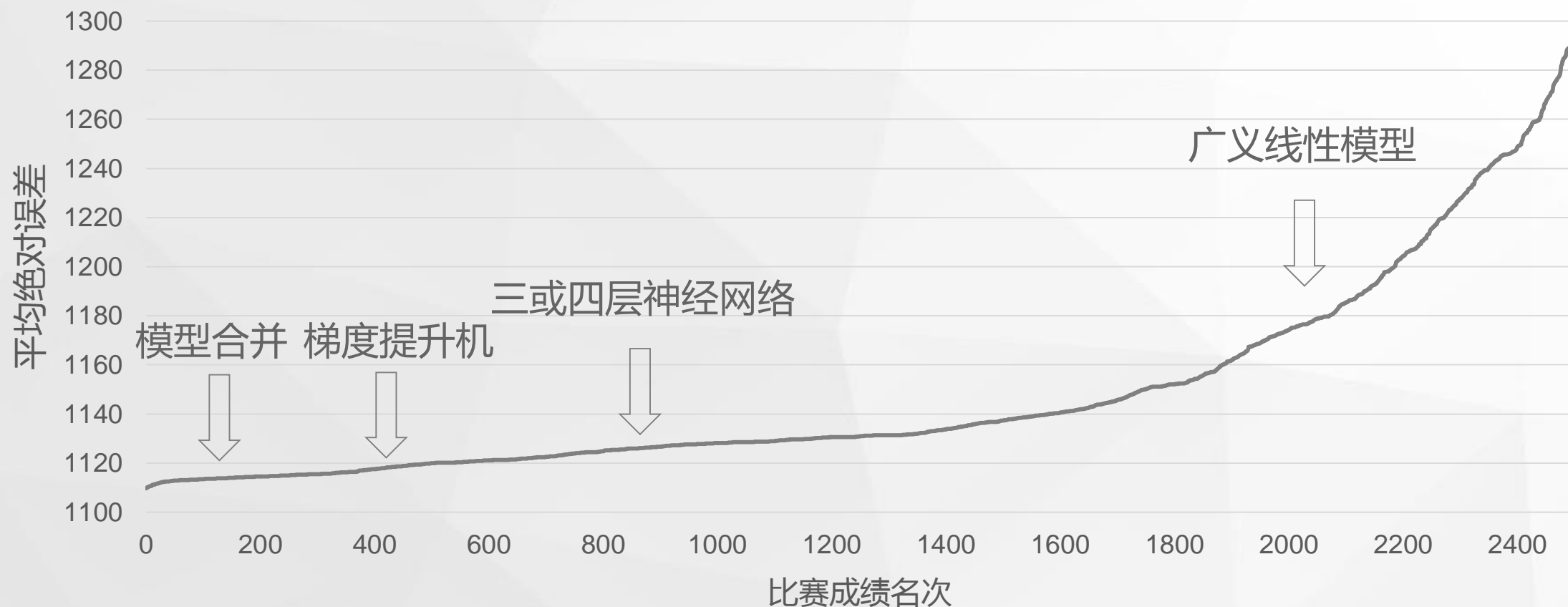


▶ 目录

- ▶ 机器学习简介
- ▶ 机器学习在保险精算领域的应用
- ▶ 机器学习模型性能比较
- ▶ 小结与问答环节

案例一：美国Allstate保险公司索赔强度预测建模竞赛

主要模型在Allstate竞赛中的性能



案例二：广义线性模型 vs. 梯度提升机模型 - 问题和建模方法描述

问题：数据源是10万保单的定价因子和相应的出险记录；除了基本的数据清理外，没有对数据进行任何处理

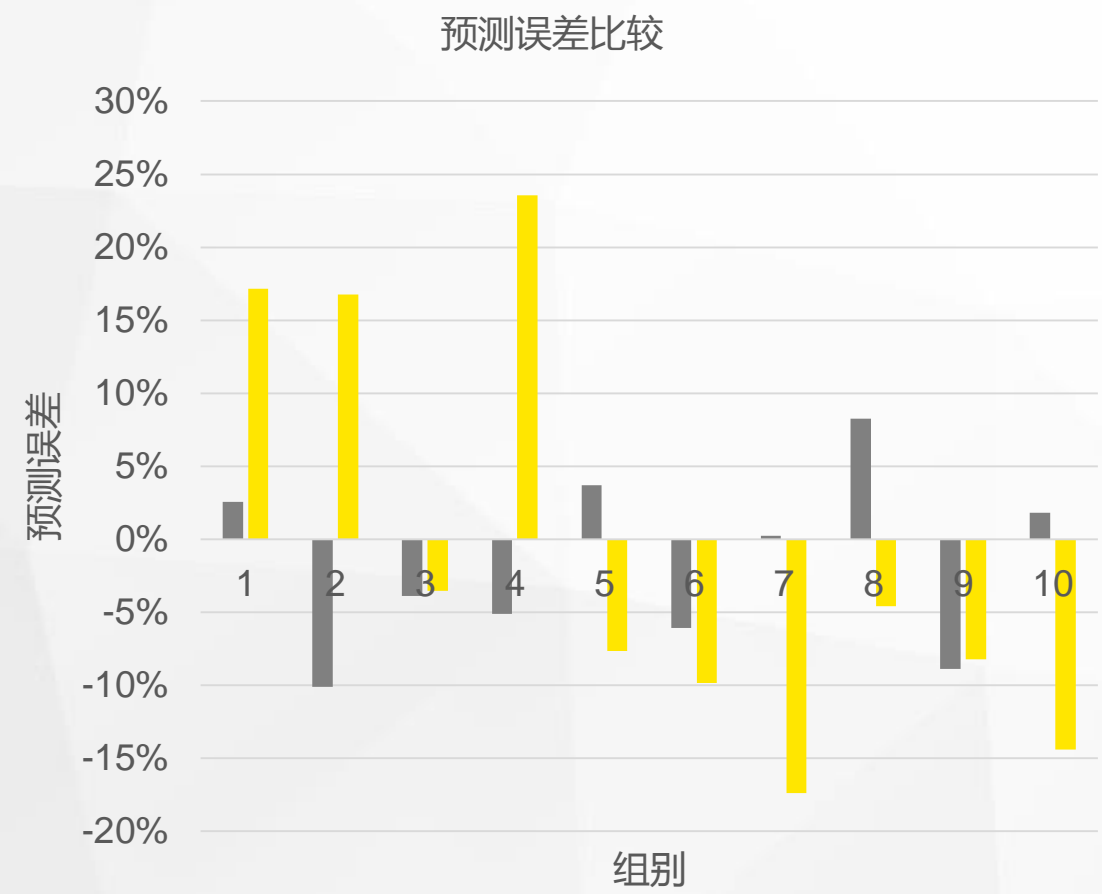
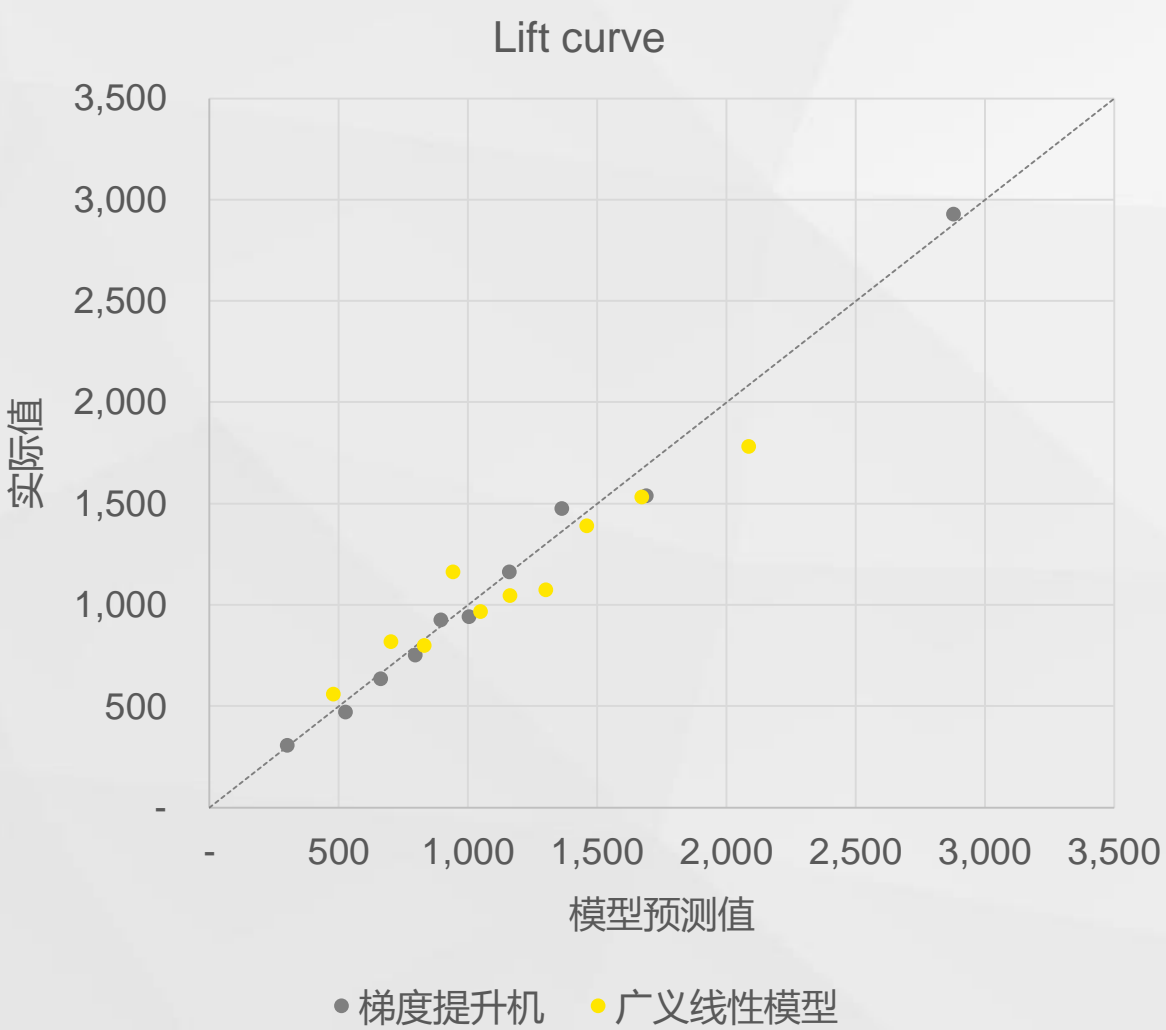
随机抽样8万用于模型建立；另外2万用于模型性能的比较

广义线性模型方法：SAS建模；传统频率和赔付强度模型；主要使用p-value决定因子的取舍

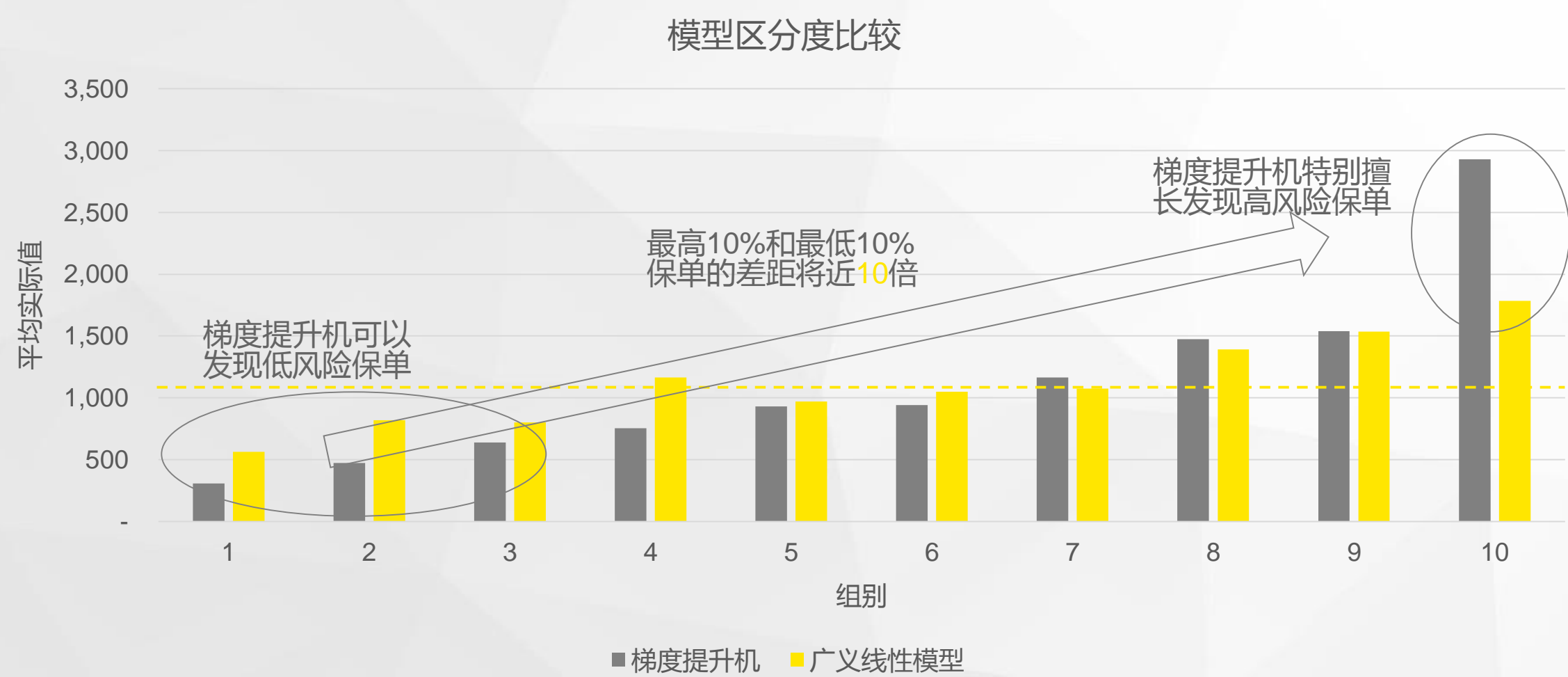
梯度提升机模型方法：Python建模；8万的80%（即6.4万训练数据），剩下的20%用于超参数的选择（验证数据）；直接对赔付金额建模

主要使用Lift curve比较模型性能：根据模型预测值从低到高排序，然后按照排序结果将保单平均分成10组，计算每组的预测值和实际值

案例二：广义线性模型 vs. 梯度提升机模型 - 预测准确性

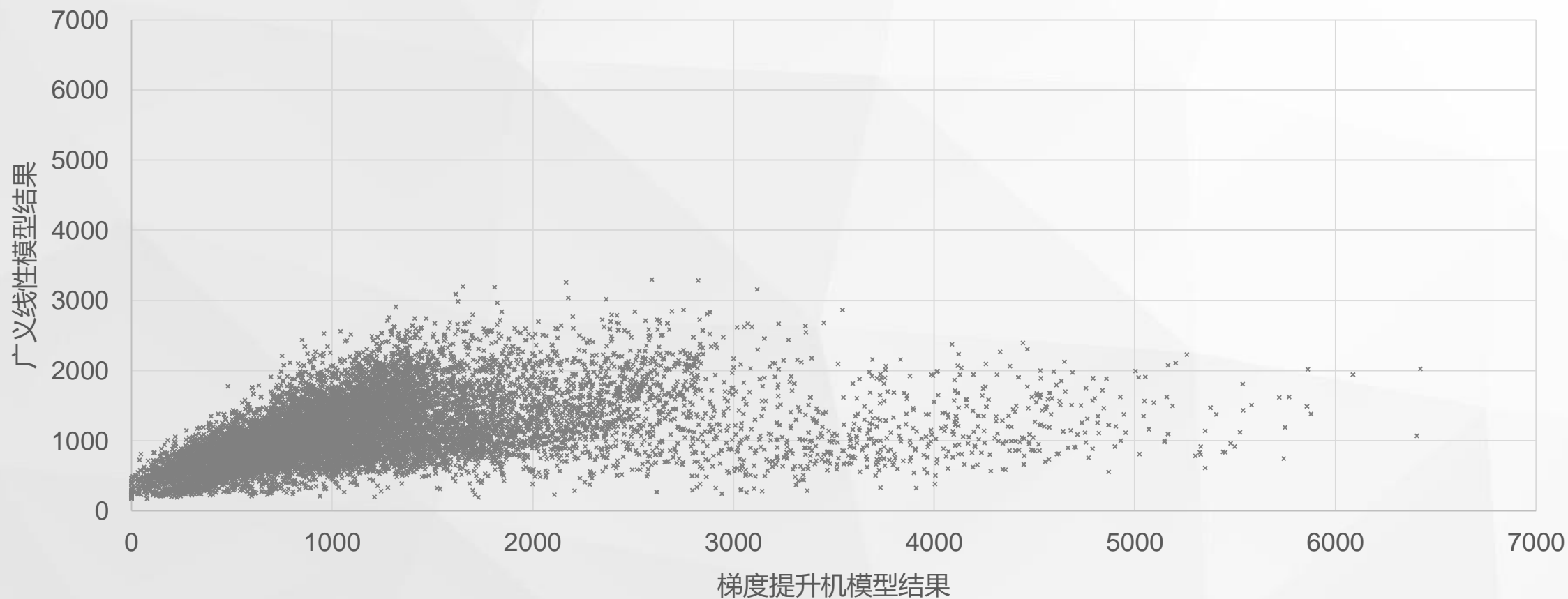


案例二：广义线性模型 vs. 梯度提升机模型 - 区分度



案例二：广义线性模型 vs. 梯度提升机模型 - 预测结果范围

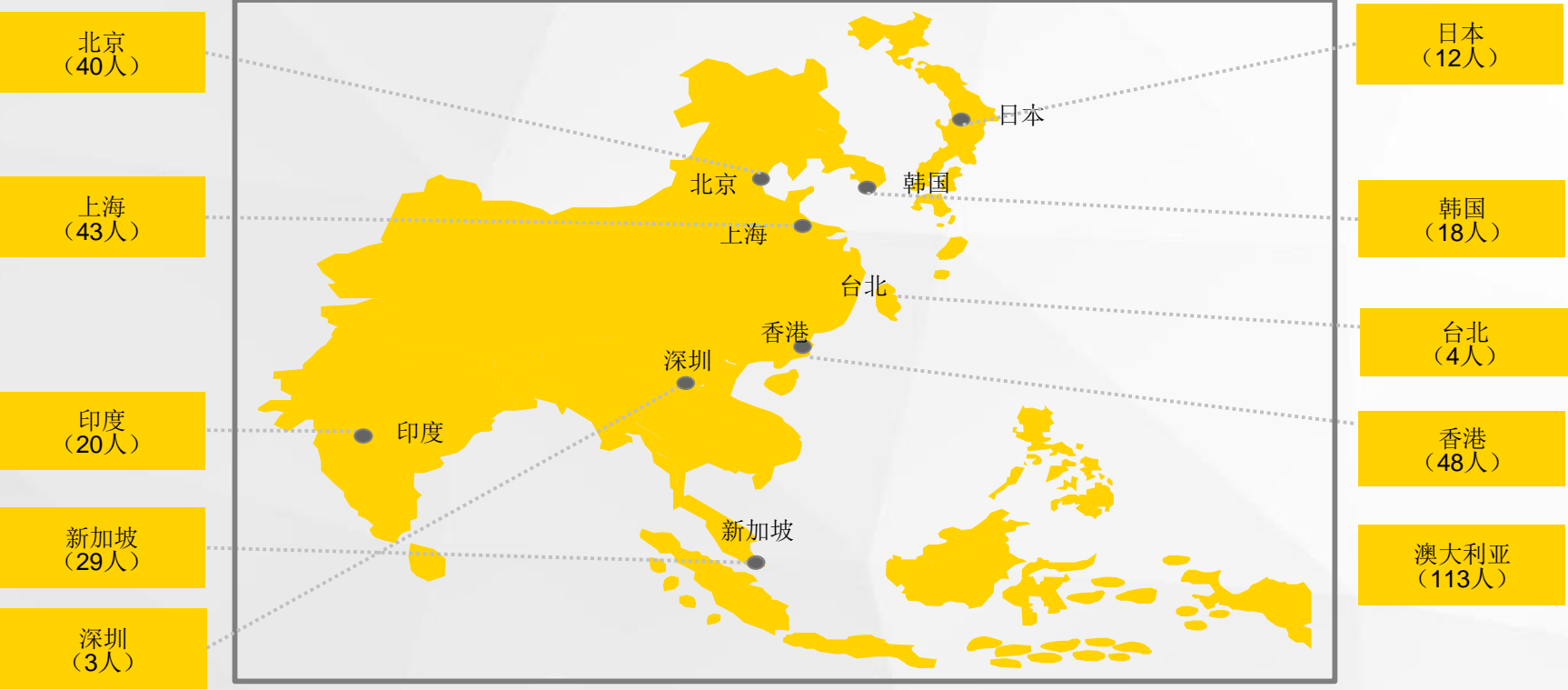
广义线性模型和梯度提升机模型结果比较



▶ 目录

- ▶ 机器学习简介
- ▶ 机器学习在保险精算领域的应用
- ▶ 机器学习模型性能比较
- ▶ 小结与问答环节

安永具有丰富经验的保险精算和风险管理咨询团队



- ▶ 安永精算咨询服务团队是亚洲最大的精算服务团队之一。通过我们在香港、上海、北京、韩国（首尔）、印度、新加坡、澳大利亚（悉尼与墨尔本）等地总共接近300人的精算团队，我们可以为这些地区提供专业的商业解决方案。
- ▶ 服务范围涵盖：车险定价咨询、外聘精算师、第三方准备金审核、经济资本模型、市场准入分析、大数据分析、精算模型开发等等众多领域。通过有效地整合及协调内部资源，我们为市场提供最优质和专业的咨询服务。

安永保险数据挖掘与建模主要联系人



姚佶 博士

电话: +86 173 1560 0803
Email: Jeff.Yao@cn.ey.com

- ▶ 现任安永(中国)保险行业与精算服务总监, 兼任上海交通大学中英精算项目学术总监, 在保险精算行业有超过十二年的工作经验, 致力于数据在保险领域的理论和实践应用
- ▶ 联合编著《预测模型在精算科学中的应用》一书
- ▶ 在加入安永(中国)之前, 他在英国致力于机器学习和大数据挖掘的全新软件Tyche的开发, 是该软件的创始团队成员之一
- ▶ 除日常工作外, 他现任世界多所大学和国际精算机构协会的考官, 曾在英国肯特大学兼职讲授应用精算学硕士课程多年。他是英国高等教育学会会士(FHEA)
- ▶ 数学与统计专业博士, 英国精算师(FIA), 特许企业风险分析师(CERA)

The background of the entire page is a close-up, artistic photograph of a dense bundle of fiber optic cables. The cables are illuminated from below, creating a bright, glowing effect with many small, out-of-focus light points and streaks of light. The overall color palette is warm, dominated by oranges, yellows, and browns, giving it a high-tech yet organic feel.

EY 安永 | Assurance 审计 | Tax 税务 | Transactions 财务交易 | Advisory 咨询

关于安永

安永是全球领先的审计、税务、财务交易和咨询服务机构之一。我们的深刻洞察和优质服务有助全球各地资本市场和经济体建立信任和信心。我们致力培养杰出领导人才，通过团队协作落实我们对所有利益关联方的坚定承诺。因此，我们在为员工、客户及社会各界建设更美好的商业世界的过程中担当重要角色。

安永是指 Ernst & Young Global Limited 的全球组织，也可指其一家或以上的成员机构，各成员机构都是独立的法人实体。Ernst & Young Global Limited 是英国一家担保有限公司，并不向客户提供服务。如欲进一步了解安永，请浏览 www.ey.com。

© 2017 安永，中国
版权所有。

APAC no.
ED MMY

本材料是为提供一般信息的用途编制，并非旨在成为可依赖的会计、税务或其他专业意见。请向您的顾问获取具体意见。

www.ey.com/china