数据分析与可视化

Fred

预习

第一部分：可视化基础方法练习

主要参考：https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/

预习方法：仔细阅读如上链接内容，课程会着重讲解以上内容中的重点的方案，后续课件有，预习时可以自己将网页的代码照着敲一遍

本次课程不会单独提供额外的 ipynb 文件，所以代码都在参考资料中有。

第二部分：共享单车项目练习

主要参考：https://www.kaggle.com/viveksrinivasan/eda-ensemble-model-top-10-percentile

预习方法：仔细阅读如上链接内容的数据分析部分，学有余力的同学可以研究下模型预估部分，课程上会讲解下代码，预习时可以自己将网页的代码照着敲一遍

本次课程不会单独提供额外的 ipynb 文件，所以代码都在参考资料中有

数据可视化

为什么需要数据可视化

答疑解惑，洞悉灼见，人人皆宜

真知灼见一目了然

查看图片比逐行逐列地阅读数字更有助于人们理解数据。通过可视化数据，您可以更有效地提出和回答各种重要问题：

- 哪些地区的销售在增长
- 推动增长的因素是什么
- 接受不同服务的客户，都有哪些特点

把可视化分析与数据科学的力量相结合，让您的数据从利用不足的资产，摇身一变成为竞争优势。

[1] https://www.tableau.com/zh-cn/resource/data-visualization#

数据可视化目标

- Correlation
- Deviation
- Ranking
- Distribution
- Composition
- Change
- Groups

[1] https://www.machinelearningplus.com/plots/top-50-matplotlib-visualizations-the-master-plots-python/

练习

Correlation

- Scatter plot with line of best fit
- Correllogram

Deviation

- Diverging Bars
- Area Chart

Ranking

- Ordered Bar Chart
- Slope Chart

Distribution

- Density Curves with Histogram
- Box Plot

Composition

- Waffle Chart
- Pie Chart

Change

- Time Series with Peaks and Troughs Annotated
- Plotting with different scales using secondary Y axis

Groups

- Andrews Curve
- Cluster Plot

Kaggle Bike Sharing Demand 案例分析

背景

Bike sharing systems are a means of renting bicycles where the process of obtaining membership, rental, and bike return is automated via a network of kiosk locations throughout a city. Using these systems, people are able rent a bike from a one location and return it to a different place on an as-needed basis. Currently, there are over 500 bike-sharing programs around the world.

背景 (续)

The data generated by these systems makes them attractive for researchers because the duration of travel, departure location, arrival location, and time elapsed is explicitly recorded. Bike sharing systems therefore function as a sensor network, which can be used for studying mobility in a city. In this competition, participants are asked to combine historical usage patterns with weather data in order to forecast bike rental demand in the Capital Bikeshare program in Washington, D.C.

Google 翻译中文版

自行车共享系统是租赁自行车的一种方式，其中通过整个城市的自助服务终端网络自动获得会员资格，租赁和自行车返回的过程。使用这些系统，人们可以从一个地方租用自行车，并根据需要将其返回到不同的地方。目前，全世界有超过 500 个自行车共享计划。

这些系统生成的数据使其对研究人员具有吸引力，因为明确记录了旅行的持续时间，出发地点，到达地点和经过的时间。因此，自行车共享系统用作传感器网络，其可用于研究城市中的移动性。在本次比赛中，参与者被要求将历史使用模式与天气数据相结合，以预测华盛顿特区 Capital Bikeshare 计划中的自行车租赁需求。

评估指标

the Root Mean Squared Logarithmic Error (RMSLE)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

参考解法

https://www.kaggle.com/viveksrinivasan/eda-ensemble-model-top-10-percentile