2022-3-17 Meta 买它 infra research data scientist 电面

ML/STATISTICS: credit fraud,

Q1: given amount and distance as features, what algorithm you will use?

Answer: Build a classification model to predict probability of fraud.

Q2: what other algorithms you can think of and what are the pro and cons compared to the one you proposed in Q1.

Answer:
2 features -> desicion tree/boosting/deep learning is not adequate.

Decision Tree:

* Not be efficient because lots of data but very few features

KNN:

* Frauds change over time, not a good patterns as new tech used in the new fraud cases
* Save all the data but not training needed

Anormaly Dection (to be reviewed):

* Distribution of individual features

Logistic regression:

* Good interpretability
* Score fast
* Training is relatively slow

Q3: coefficient of amount to fraudulence if 0.10 with standard error 0.02, what's the relationship between amount and fraudulence? Is it statistically significant? How do you prove it?

Answer: (See ESL Page 124) Each unit increase in the distance accounts for an increase in the odds of fraudulence of exp(0.10)~=1.105 or 10.5%. The Z score is 0.10/0.02=5 which means the coeffient is significant. The is proved by the CLT.