Week 2

How to Win a Data Science Competition: Learn from Top Kagglers

Exploratory Data Analysis

We will start this week with Exploratory Data Analysis (EDA). It is a very broad and exciting topic and an essential component of solving process. Besides regular videos you will find a walk through EDA process for Springleaf competition data and an example of prolific EDA for NumerAI competition with extraordinary findings.

Less

Key Concepts

- Describe the major visualization tools
- Generate hypotheses about data
- Inspect the data and find golden features
- Examine and analyze various plots and other data visualizations

Less

Exploratory data analysis

Reading: Week 2 overview 10 min

Video: Exploratory data analysis 7 min Resume

Video: Building intuition about the data 6 min

Notebook: Reading material for video 2 20 min

Video: Exploring anonymized data 15 min

Notebook: Notebook for video 3 screencast

Video: Visualizations 11 min

Video: Dataset cleaning and other things to check 7 min

Quiz: Exploratory data analysis 4 questions Due Oct 5, 1:59 AM CDT

Reading: Additional material and links 10 min

EDA examples

Notebook: Notebook for the screencast

Video: Springleaf competition EDA I 8 min

Video: Springleaf competition EDA II 16 min

Video: Numerai competition EDA 6 min

Validation

In this module we will discuss various validation strategies. We will see that the strategy we choose depends on the competition setup and that correct validation scheme is one

we choose depends on the competition setup and that correct validation scheme is one of the bricks for any winning solution.

Key Concepts

- Describe validation process and its purpose
- Compare validation strategies
- Identify train/test split in a competition
- Identify and analyze validation problems

Less

Validation

Video: Validation and overfitting 9 min Resume

Video: Validation strategies 7 min

Reading: Validation strategies 10 min

Video: Data splitting strategies 14 min

Video: Problems occurring during validation 20 min

Practice Quiz: Validation 4 questions

Quiz: Validation 4 questions Due Oct 5, 1:59 AM CDT

Reading: Comments on quiz 10 min

Reading: Additional material and links 10 min

Data Leakages

Finally, in this module we will cover something very unique to data science competitions. That is, we will see examples how it is sometimes possible to get a top position in a competition with a very little machine learning, just by exploiting a data leakage.

Less

Key Concepts

- Embrace the concept of data leakage
- Find and exploit typical data leakages
- Probe public leaderboard

Less

Data leakages

Video: Basic data leaks 6 min Resume

Video: Leaderboard probing and examples of rare data leaks 9 min

Video: Expedia challenge 9 min

Quiz: Data leakages 4 questions Due Oct 5, 1:59 AM CDT

Reading: Comments on quiz 10 min

Notebook: Data leakages

Programming Assignment: Data leakages 3h Due Oct 5, 1:59 AM CDT

Peer-graded Assignment: Data leakages 30 min Due Oct 5, 1:59 AM CDT

leakages

Review Your Peers: Data leakages Due Oct 8, 1:59 AM CDT

Reading: Additional material and links 10 min

Reading: Final project advice #2 10 min

Discussion Prompt: Looking for a team 5 min