Competitive Machine Learning

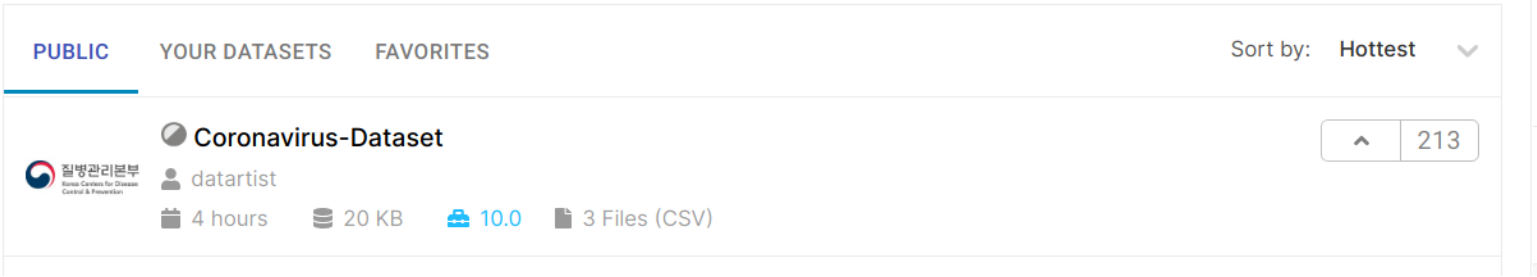
Workshop — March 22, 2021

Competing with Quantitative Sciences

- I want to call your attention to and/or provide insight into how methods/techniques you learn in various COSC, STAT, MATH and DATA courses can be immediately used in a variety of online competitions
- Some of this will definitely come with disclaimers!
- I will break this into a couple parts
  1. Kaggle: introduce, provide some guidance and advice
  2. Daily Fantasy Sports: introduce, provide guidance, walkthrough an example

Kaggle — What is it?

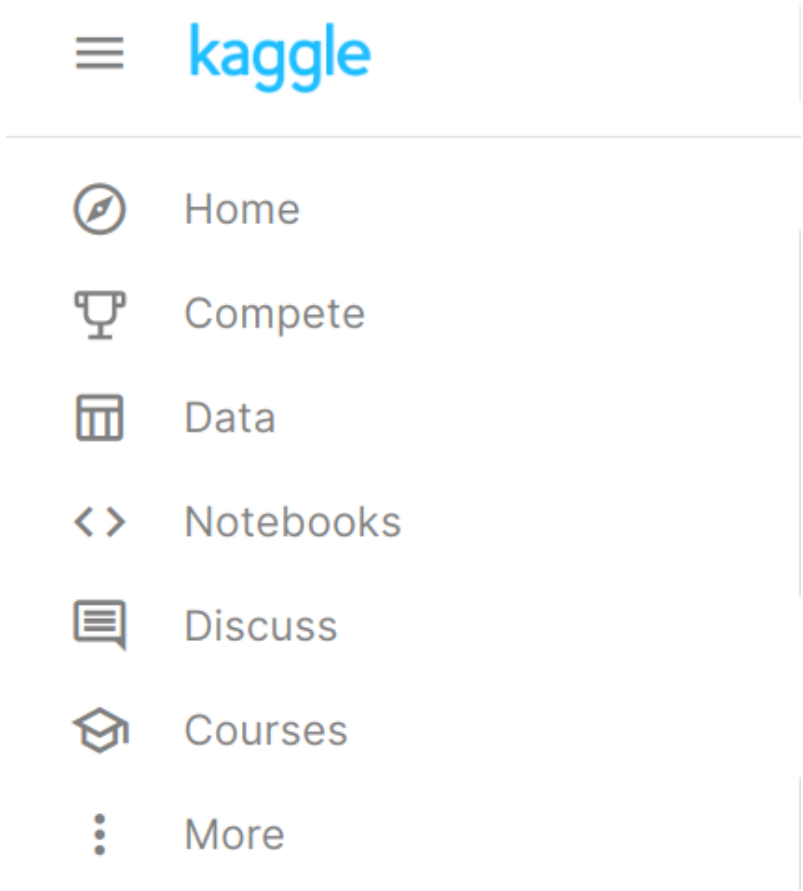
- Kaggle.com is an online data science competition hosting platform
- Beyond competitions, it also serves as an excellent repository for interesting data sets.



- Disclaimer: I have not competed in any kaggle competitions, though I have sourced many data sets from there
- My advice therefore will be a mix of my general modelling experiences (consulting, research, etc), combined with other sources on how kaggle competitions operate.

Kaggle — What is it?

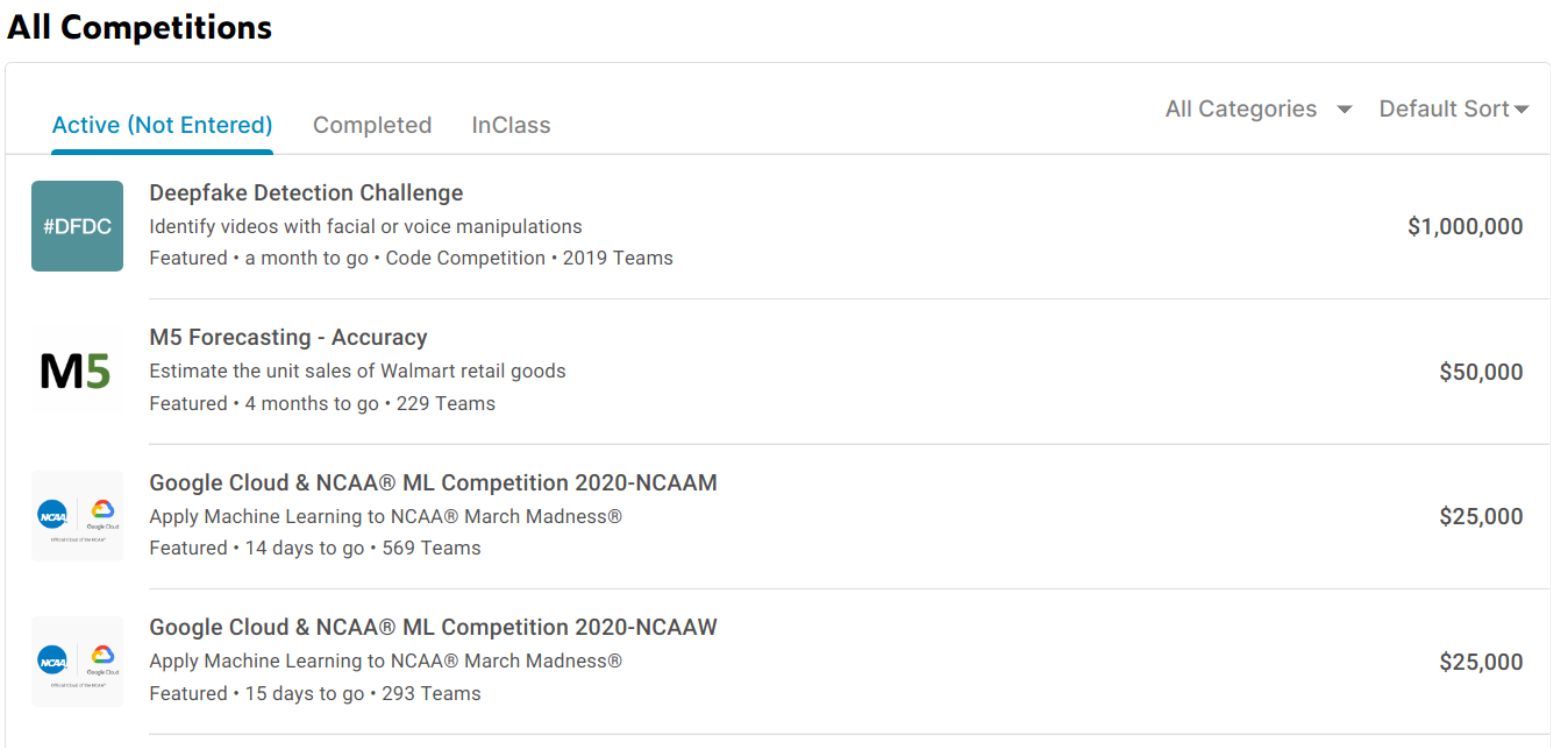
- Alright, some light exploration



- First up, compete...

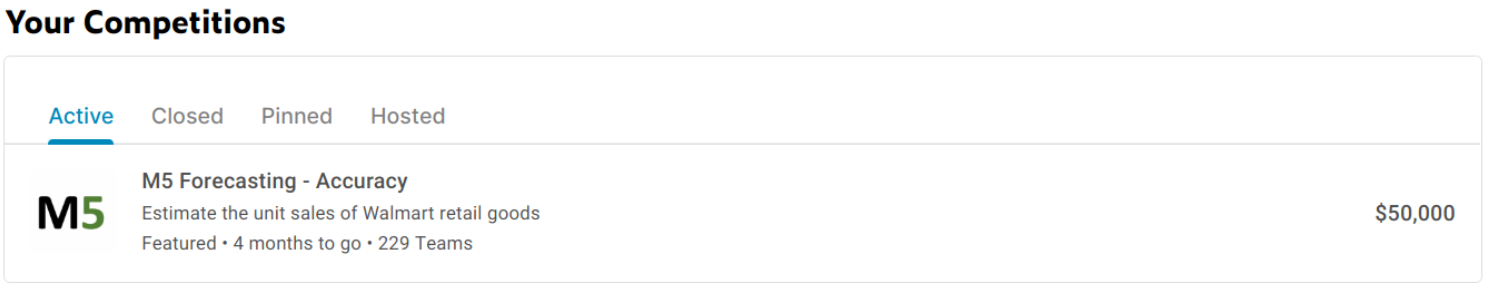
Kaggle — Competitions

- This page summarizes some competitions that we're running last year (website framework hasn't changed much).



Kaggle — Competitions

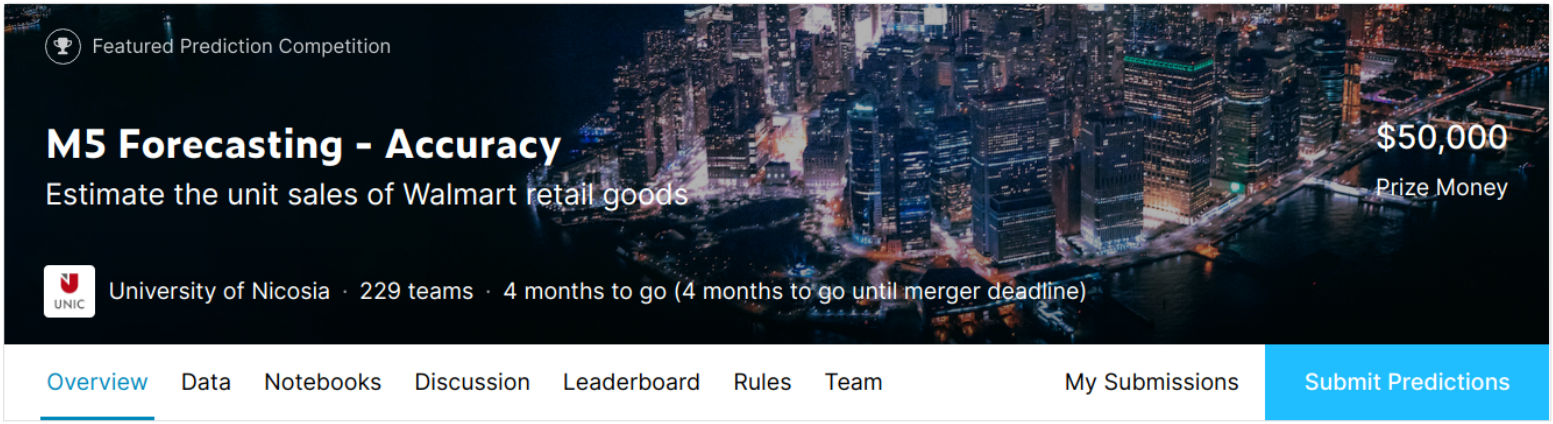
- From trying to take a peak at some data sets, I've already 'entered' a competition.



- Let's take a closer look at the M5 Forecasting - Accuracy competition

Kaggle — Competitions

- The first page will land you on an overview...starting with a description



Overview

Description

Evaluation

Timeline

Prizes

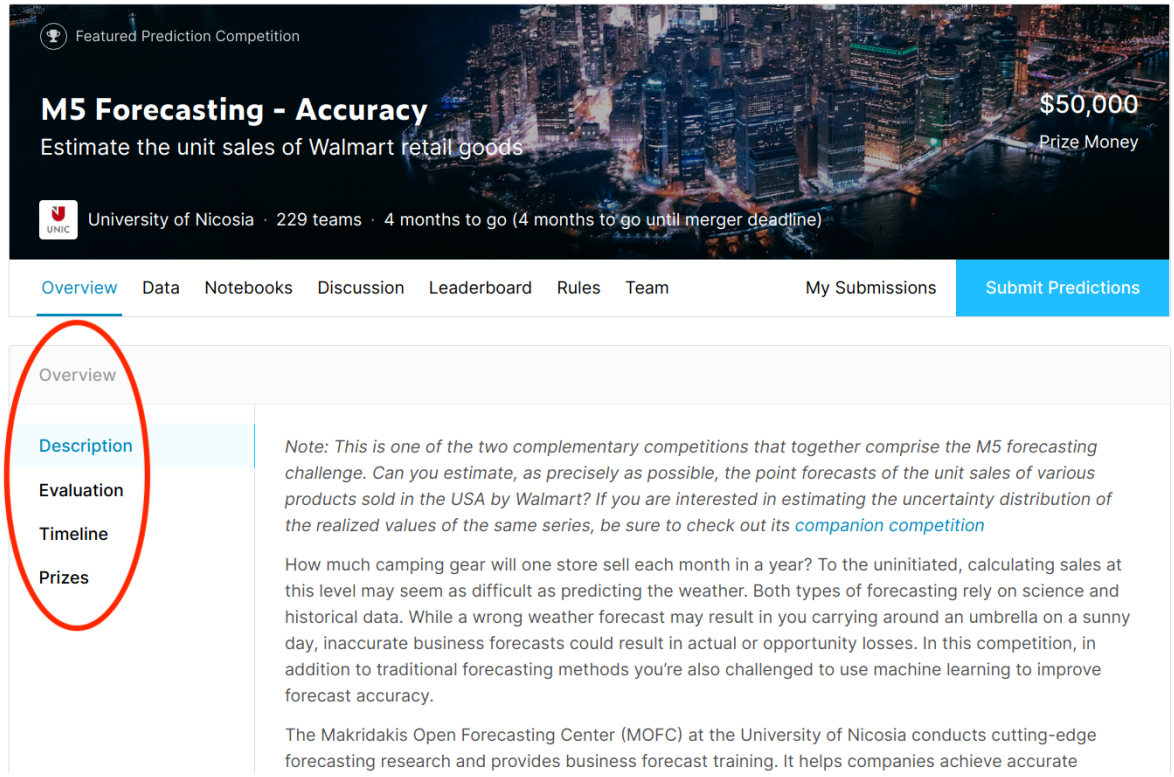
Note: This is one of the two complementary competitions that together comprise the M5 forecasting challenge. Can you estimate, as precisely as possible, the point forecasts of the unit sales of various products sold in the USA by Walmart? If you are interested in estimating the uncertainty distribution of the realized values of the same series, be sure to check out its companion competition

How much camping gear will one store sell each month in a year? To the uninitiated, calculating sales at this level may seem as difficult as predicting the weather. Both types of forecasting rely on science and historical data. While a wrong weather forecast may result in you carrying around an umbrella on a sunny day, inaccurate business forecasts could result in actual or opportunity losses. In this competition, in addition to traditional forecasting methods you're also challenged to use machine learning to improve forecast accuracy.

The Makridakis Open Forecasting Center (MOFC) at the University of Nicosia conducts cutting-edge forecasting research and provides business forecast training. It helps companies achieve accurate

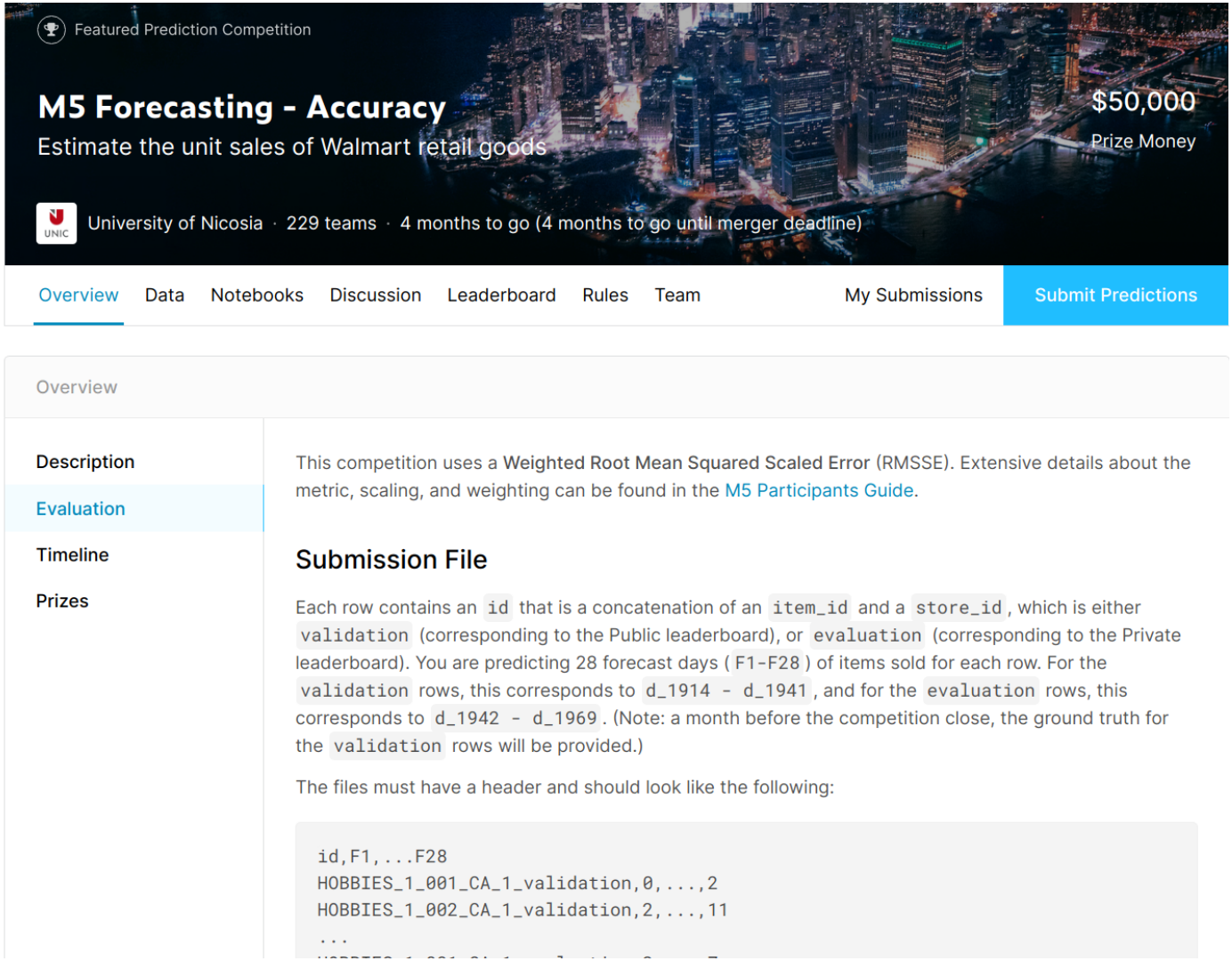
Kaggle — Competitions

- We'll look along the left-hand menu first



Kaggle — Competitions

- Next in the overview is how the competition will be evaluated



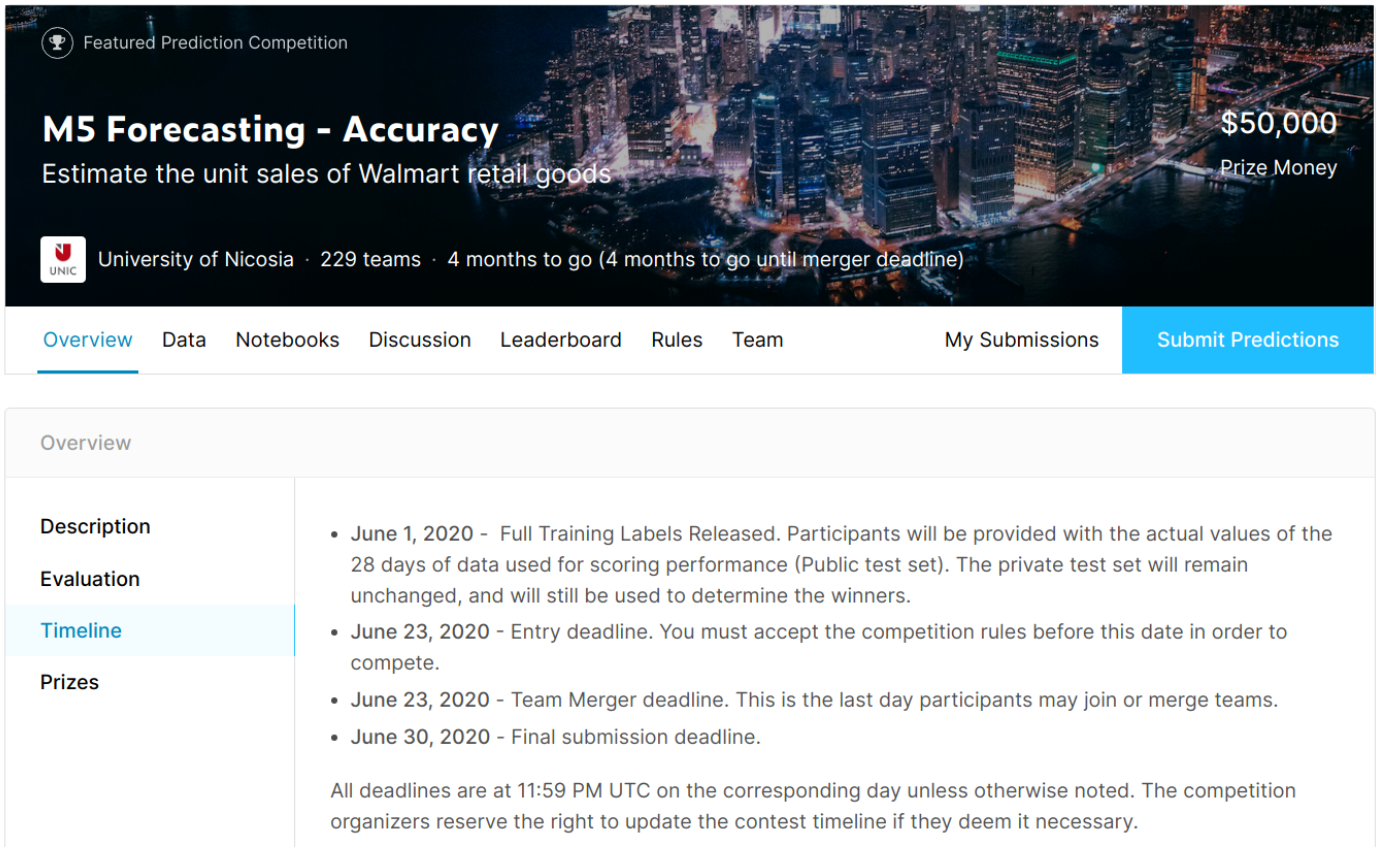
Comments on Evaluation

- This is one of the most important pieces of information to consider for the competition
- Competitions can be scored in several manners, and the choice of scoring can determine what your modeling concerns are.
- For example, scoring on minimizing squared error vs minimizing absolute error.
- The former would suggest avoiding any really large errors, the latter would suggest avoiding making lots of medium errors...

	Mod1	Mod2	Truth
	0	9	10
	50	20	40
	90	101	100
Sum Sq Error	300	402	
Sum Abs Error	30	22	

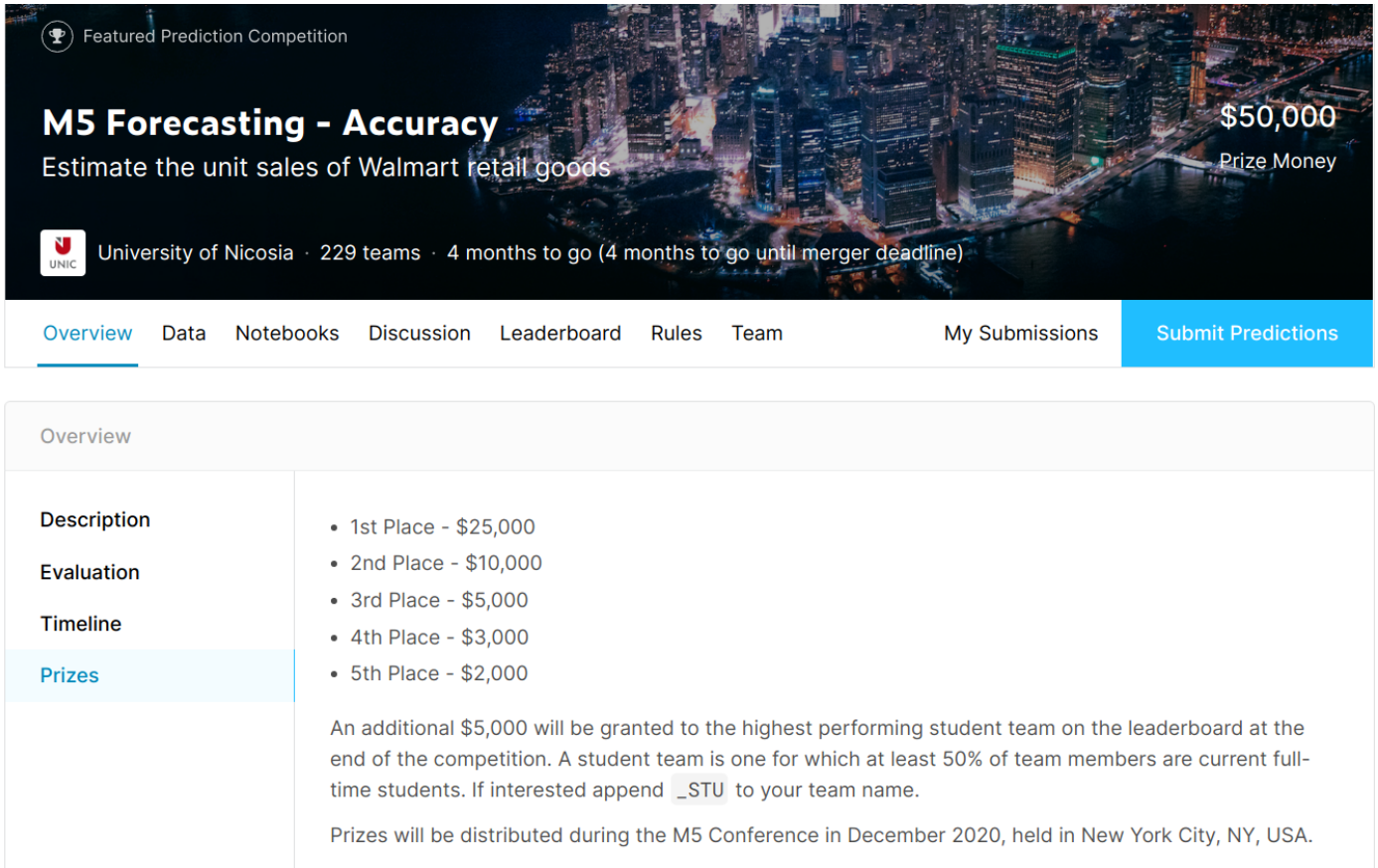
Kaggle — Competitions

- Then, a timeline for the competition



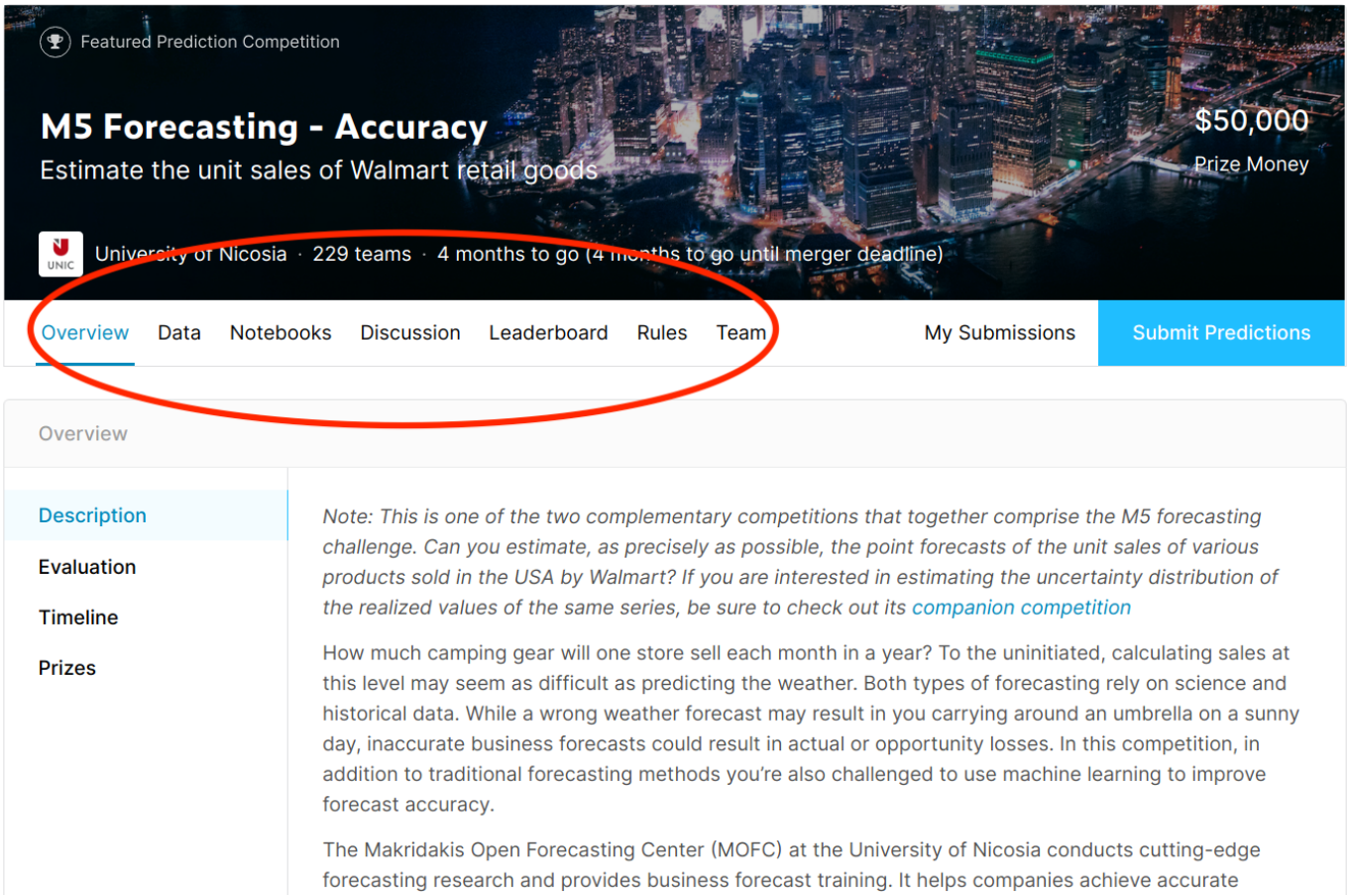
Kaggle — Competitions

- And the all-important prizes! Note that these are nothing to sniff at (though I would argue against quitting school to do this full time).



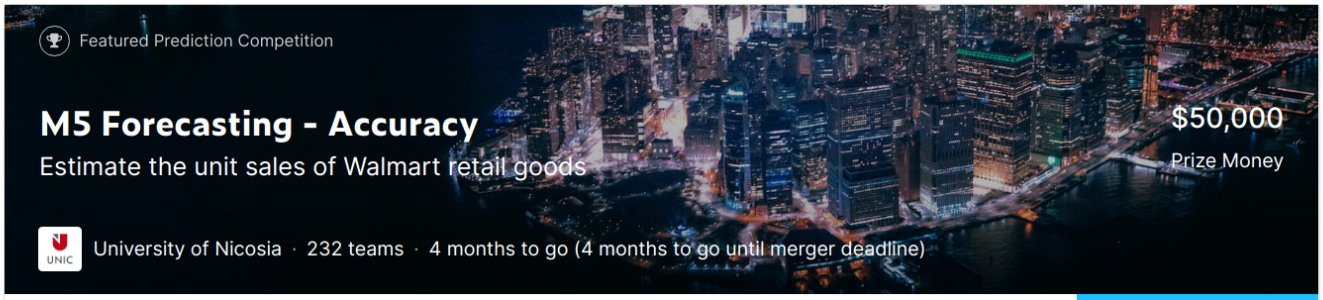
Kaggle — Competitions

- Now let's explore the top menu



Kaggle — Competitions

- Data will provide a more detailed description...and the data. You can generally preview the data here, and also download it.



Overview | Data | Notebooks | Discussion | Leaderboard | Rules | Team | My Submissions | Submit Predictions

Data Description

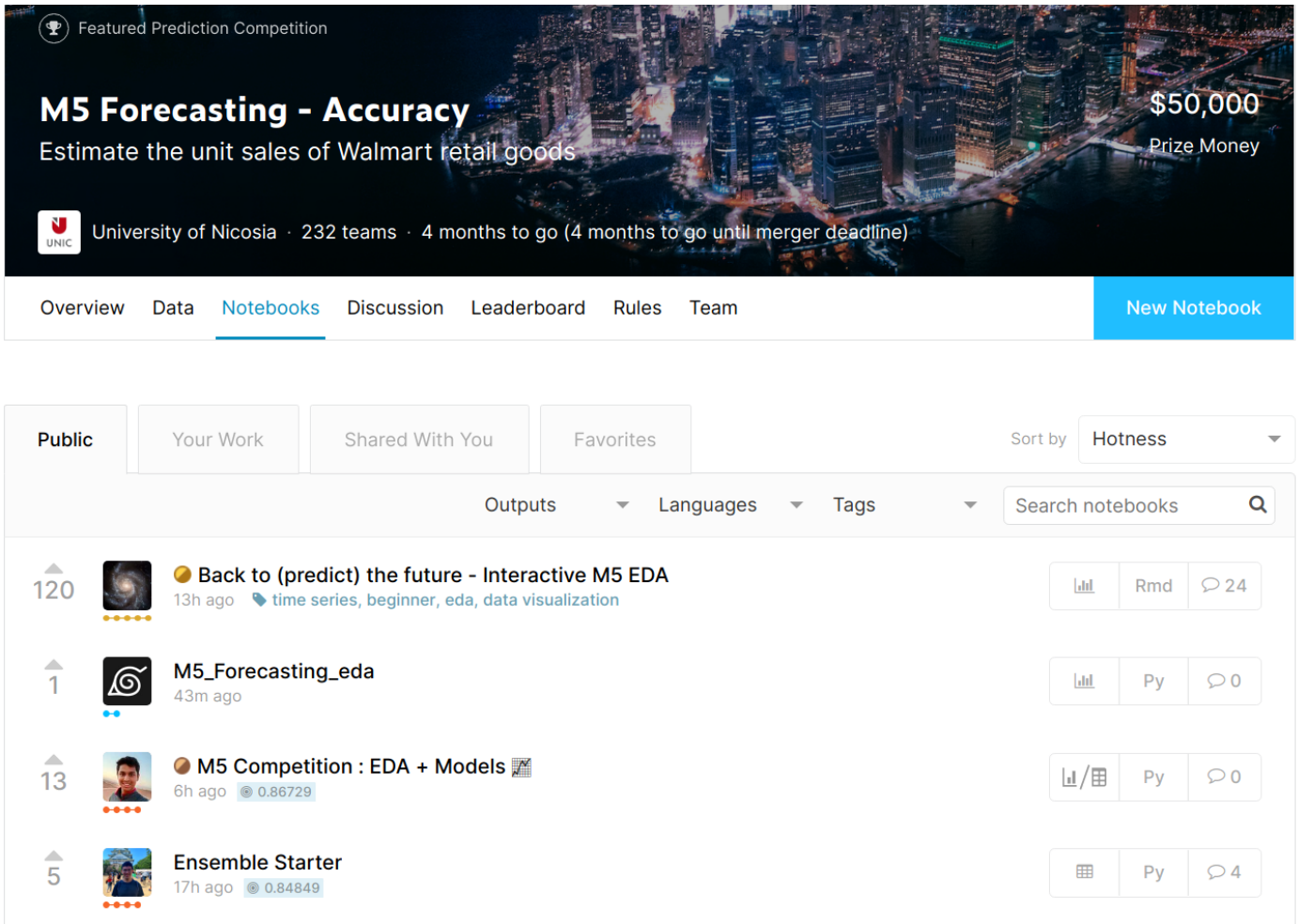
In the challenge, you are predicting item sales at stores in various locations for two 28-day time periods. Information about the data is found in the M5 Participants Guide.

Files

- `calendar.csv` - Contains information about the dates on which the products are sold.
- `sales_train_validation.csv` - Contains the historical daily unit sales data per product and store `[d_1 - d_1913]`
- `sample_submission.csv` - The correct format for submissions. Reference the Evaluation tab for more info.
- `sell_prices.csv` - Contains information about the price of the products sold per store and date.
- ***`sales_train_evaluation.csv`*** - *Available once month before competition deadline. Will include sales* ***`[d_1 - d_1941]`***

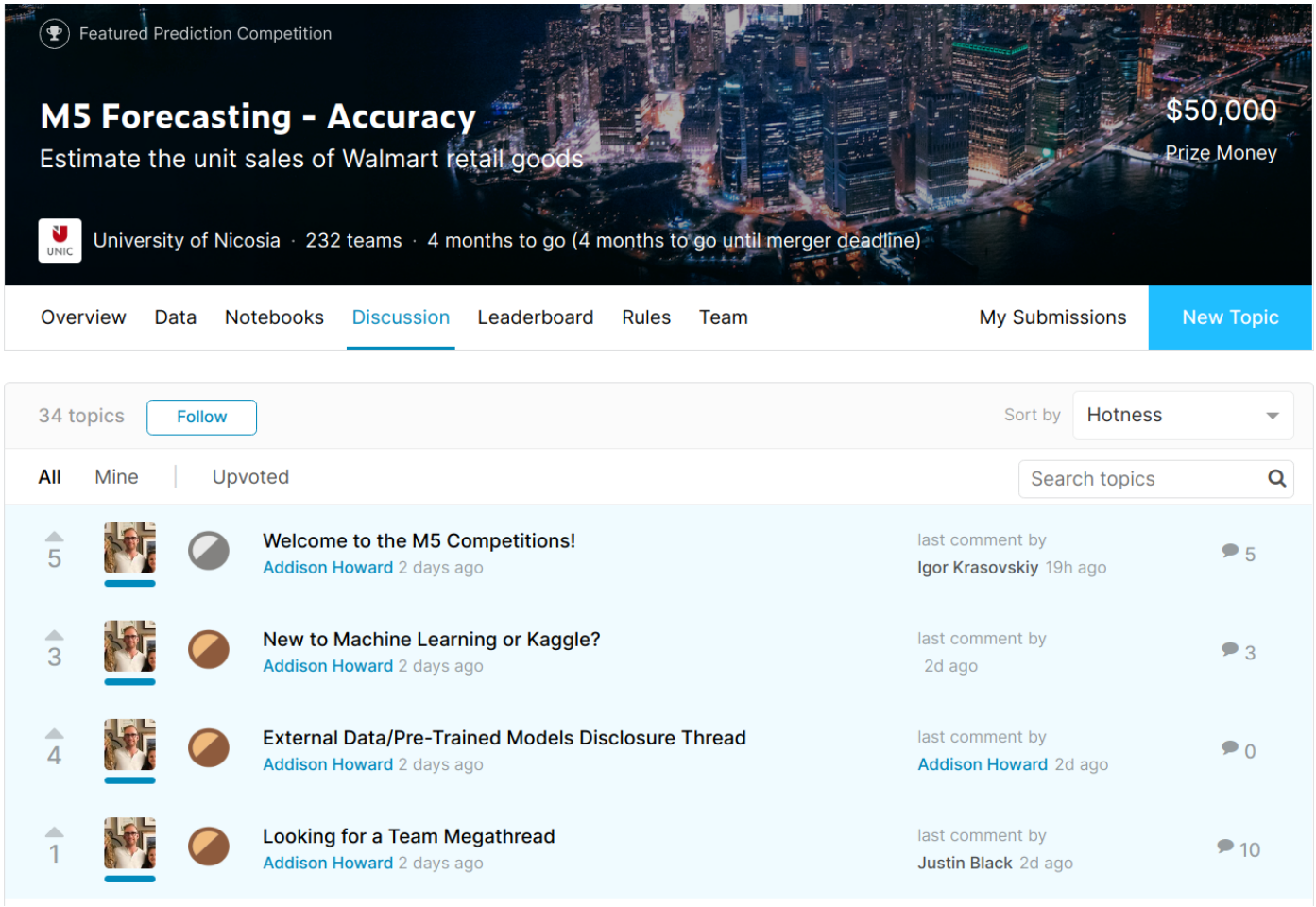
Kaggle — Competitions

- Next are notebooks...these are public notebooks of analysis on the data. These are great learning tools and will also give you some sense of the depth/expertise of your competition.



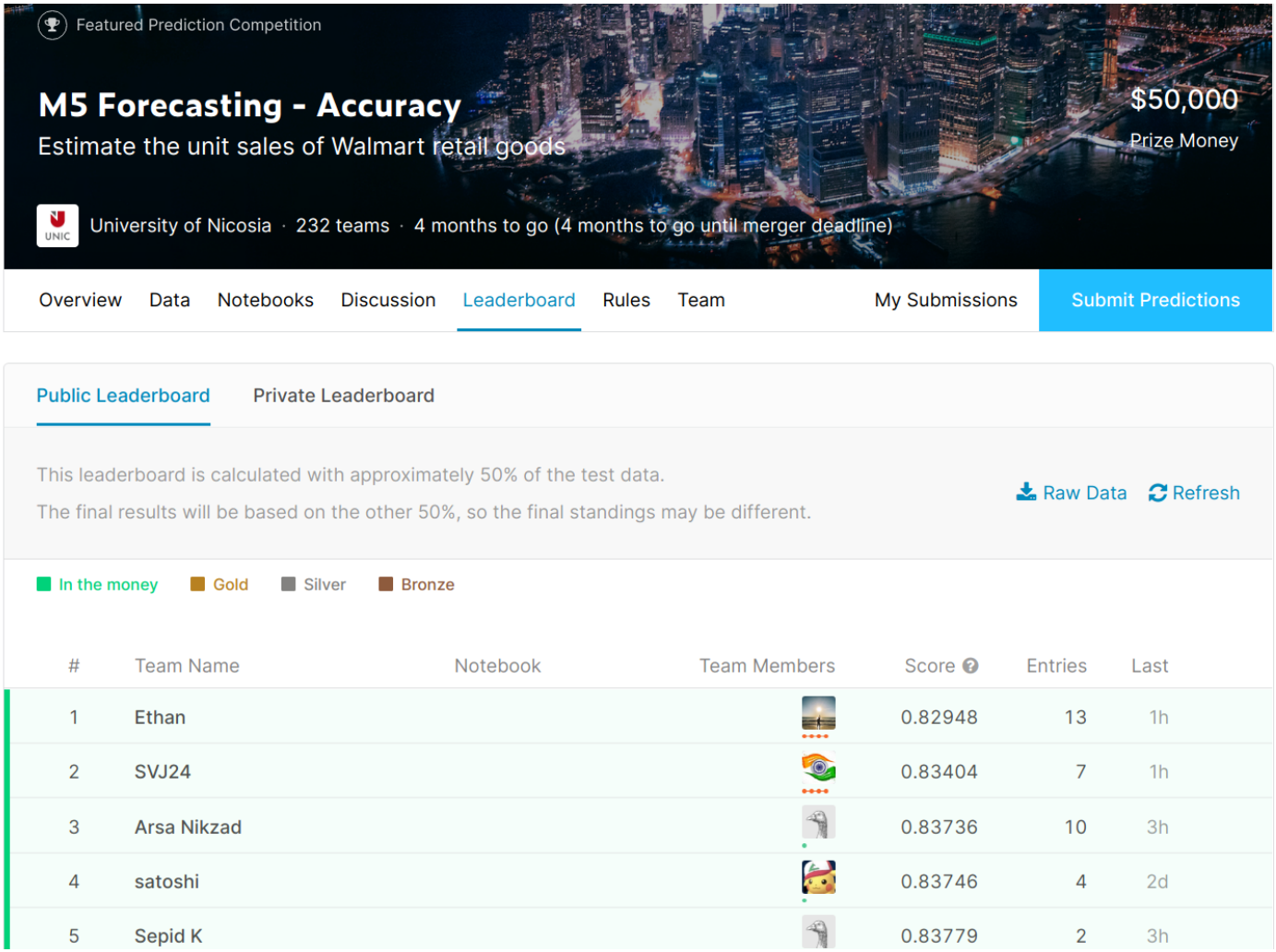
Kaggle — Competitions

- Discussion is a forum area



Kaggle — Competitions

- Leaderboards! Note that these are split into public vs private...



Primer: Training vs Testing

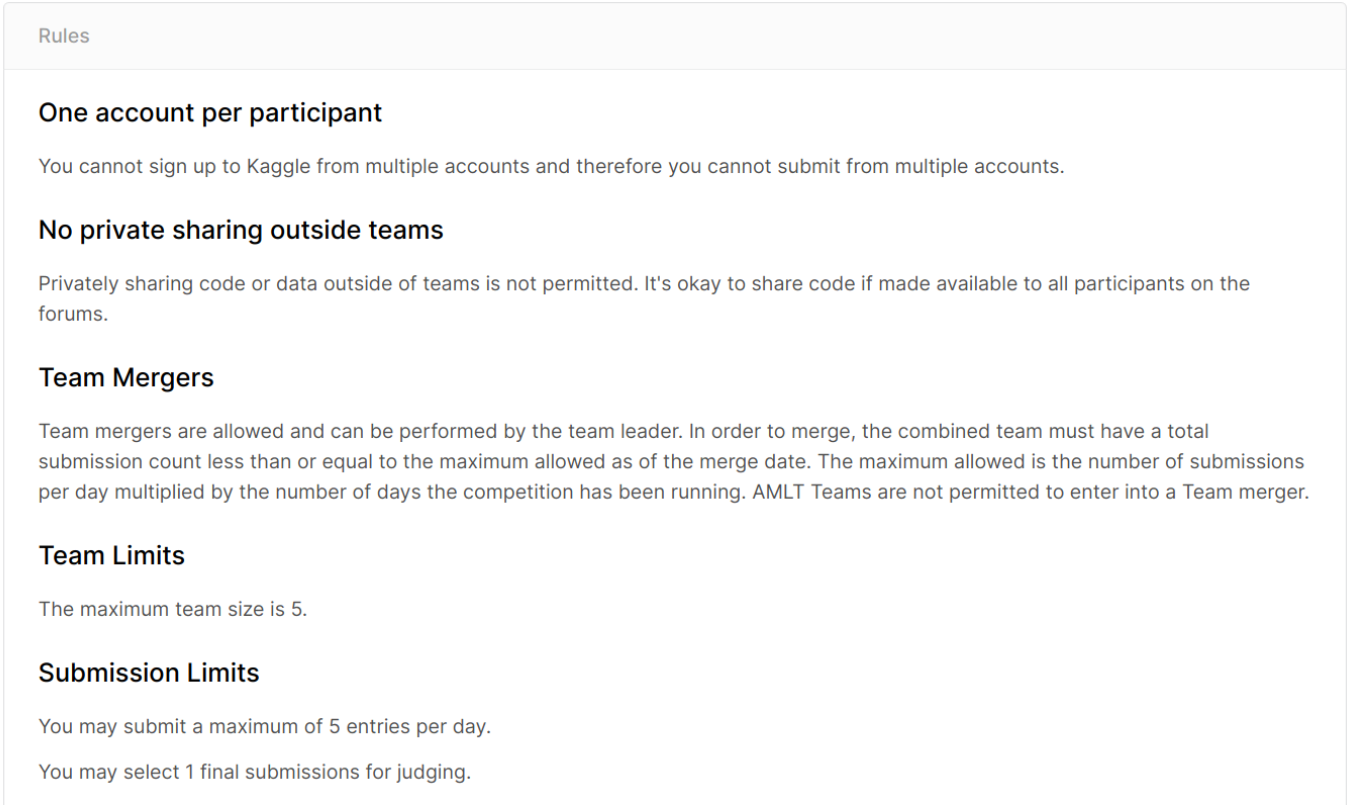
- Supervised ML methods are fitted (or trained) on observed data with a known response variable — the training data.
- The fitted model is then used to predict for some held out data (not used during fitting) to check how well the model predicts in a realistic scenario — the testing data.
- This is to help avoid the plague of overfitting.
- Kaggle further splits testing sets in order to provide feedback (public leaderboard) on a subset of the testing data, while witholding the results on the remainder of the test set (private leaderboard) until the competition closes

Something to be aware of

- Interestingly, this setup leads to a different kind of overfitting than we are often worried about — we should be cognizant of the potential of overfitting on the subset of the test set!
- While you can't explicitly train your supervised model on the test set used for that leaderboard, you are provided feedback on it through the evaluation index.
- And naturally, you are trying to climb that leaderboard...so you will try to make modelling choices that predict better on the public test set...which may not always predict better on the private test set!
- One way to guard against this is to ensure that you see improvement on both your training data (through cross-validation) in tandem with improvement on the leaderboard — but it's not foolproof.

Kaggle — Competitions

- Details on the rules — these can change from competition to competition, and it's probably worthwhile ensuring that you're eligible to win the prize you are competing for!



Rules

One account per participant

You cannot sign up to Kaggle from multiple accounts and therefore you cannot submit from multiple accounts.

No private sharing outside teams

Privately sharing code or data outside of teams is not permitted. It's okay to share code if made available to all participants on the forums.

Team Mergers

Team mergers are allowed and can be performed by the team leader. In order to merge, the combined team must have a total submission count less than or equal to the maximum allowed as of the merge date. The maximum allowed is the number of submissions per day multiplied by the number of days the competition has been running. AMLT Teams are not permitted to enter into a Team merger.

Team Limits

The maximum team size is 5.

Submission Limits

You may submit a maximum of 5 entries per day.

You may select 1 final submissions for judging.

General Advice

- Evaluation is always (at least that I've seen) on a prediction metric of one form or another.
- This essentially means we can ignore one of our (DATA 311/MDS) core modelling tradeoffs: inference vs prediction!
- As such, the most common winning methodologies usually fall into one of the following (relatively large) bins
  - Deep neural nets
  - Ensemble techniques (often boosted models)

General Advice

- "Wait, if we basically know what methodologies will win, where does the competition take place?"
- The main game of most Kaggle competitions is finding (externally to the data) or generating (internally from the data) useful variables on which to fit the winning method.
- This is often referred to as 'feature engineering'.
- There are usually notebooks dedicated to this on most competitions. I would encourage you to browse these from old and current competitions to get a feel for it.

General Advice

- Often good feature engineering comes down to reasonable domain knowledge.
- AKA, it often helps to have at least some knowledge/understanding of the phenomena that you are trying to model.
- Admittedly, this is more important in a general modelling/consulting context, as it helps you to find simple models that perform 'well-enough' (again, inference vs prediction tradeoff).
- Now, we will get some practice feature engineering...but outside of Kaggle, in the world of daily fantasy sports.

DFS — What is it?

- Daily Fantasy Sports have exploded as a massive industry in the past decade.
- Draftkings and Fanduel are the two biggest platforms.
- Disclaimer: Legality of DFS in Canada is a gray area — they claim to be games of skill, others claim it is functionally equivalent to gambling. The truth is likely somewhere in between, but the variability in results is high enough that even with a significant predictive advantage (which you won't necessarily know you have), you can lose a substantial amount of money in the short-to-medium term.
- HUGE Personal Disclaimer: My discussion here should not be taken as encouragement for you to partake. I have played off-and-on for approx 6 years, with a total investment of around $50 (I really only compete in super low-stakes events).

DFS — What is it?

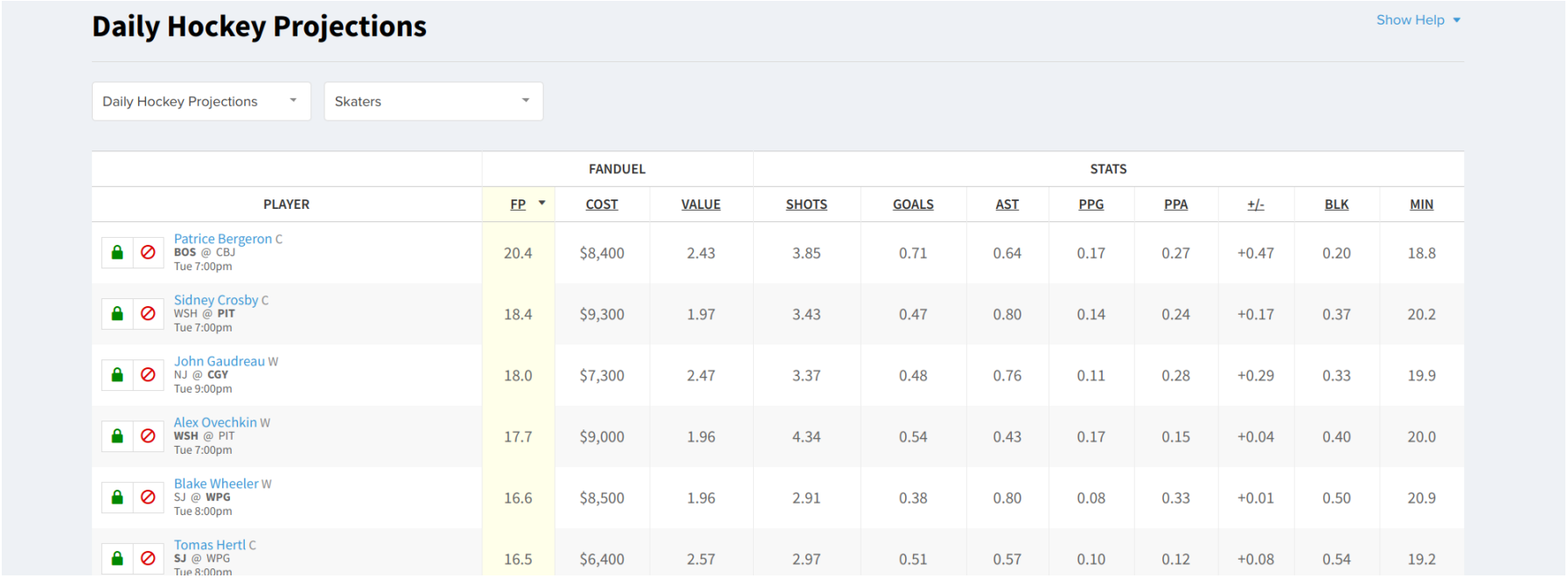
- Okay, so disclaimers out of the way...how does it work?
- That depends on the sport, competition, and website...and it's not worth our time to cover all the specifics.
- In general, you are given a certain amount of units (a salary cap) that you can use to pick players for the games that day, and better players will cost more to select.
- The players will accrue points based on good performance in their respective sport (hockey example: shots, goals, assists, blocked shots, etc) — rarer positive events are generally worth more (goals $>$ shots, etc).

DFS — Our goals

- So competing in DFS (in any sport) comes down to two basic tasks
  1. Predicting the performance of all players — specifically their total amount of fantasy points that day
  2. Selecting a team that is under budget that maximizes the sum of those predicted points
- Task 1 is a machine learning task, with a whole lot of necessity for external information and feature engineering
- Task 2 is a constrained optimization task. It can be formatted as a (binary) linear programming problem.

DFS — Our goals

- Since it's more fun (and potentially more advantageous) we will approach Task 1 from scratch
- It's worth noting however that there are several websites out there that provide their own predictions (usually using some undescribed model).



Daily Hockey Projections

Show Help

Daily Hockey Projections | Skaters

		FANDUEL			STATS							
	PLAYER	FP	COST	VALUE	SHOTS	GOALS	AST	PPG	PPA	+/-	BLK	MIN
	Patrice Bergeron C **BOS** @ CBJ Tue 7:00pm	20.4	$8,400	2.43	3.85	0.71	0.64	0.17	0.27	+0.47	0.20	18.8
	Sidney Crosby C WSH @ **PIT** Tue 7:00pm	18.4	$9,300	1.97	3.43	0.47	0.80	0.14	0.24	+0.17	0.37	20.2
	John Gaudreau W NJ @ **CGY** Tue 9:00pm	18.0	$7,300	2.47	3.37	0.48	0.76	0.11	0.28	+0.29	0.33	19.9
	Alex Ovechkin W **WSH** @ PIT Tue 7:00pm	17.7	$9,000	1.96	4.34	0.54	0.43	0.17	0.15	+0.04	0.40	20.0
	Blake Wheeler W SJ @ **WPG** Tue 8:00pm	16.6	$8,500	1.96	2.91	0.38	0.80	0.08	0.33	+0.01	0.50	20.9
	Tomas Hertl C **SJ** @ WPG Tue 8:00pm	16.5	$6,400	2.57	2.97	0.51	0.57	0.10	0.12	+0.08	0.54	19.2

DFS — Our goals

- I have built algorithms for most sports, though of varying complexity. Some I just source other website predictions (like on previous slide).
- Nowadays, I focus on golf.
- I've played golf once in the past 5 years.
- I don't (or at least didn't) enjoy watching golf on TV.
- So why golf???

DFS — Why golf?

- Golf competitions last 4 days. This has two positives from my standpoint:
  - I can field a team for 10 cents, and potentially get back 4 days of model-watching 'excitement'.
  - 4 days expands the essential 'sample size' of the competition...this reduces variability (though there is still a crazy amount of it). For example, it's very hard to predict if Connor McDavid will score a goal tonight...it's marginally easier to predict if Rory McIlroy will finish in the top 20 for an entire tournament.
- DFS Golf scoring, on the face of it, is complicated (score on each hole counts, various bonuses for birdies-in-a-row, bogey-free rounds, etc)...BUT extremely correlated with the end rank of the tournament, especially near the top of the leaderboard.
  - This means I can try to predict one response variable (tournament rank) as an effective proxy for the complicated scoring system.
  - Note: many other sports do not have this simplicity. Connor McDavid's fantasy points are pretty weakly correlated with whether, say, Edmonton wins their upcoming game.

DFS — Towards a Workshop

- Let's move towards getting our hands dirty.
- Reminder: I am not endorsing DFS in any way, shape, or form. I am also not endorsing any particular website for DFS. I am using draftkings because I am most familiar with their data setup.
- Jump to chrome (and then R...only a few parting comments on the slides for later).

DFS — Final comments

- The biggest advantages in DFS come from late-breaking news...
- Suppose Connor McDavid is announced to be sitting out 10 minutes before faceoff — since he's a strong player, a high percentage of teams (say 20-30%) are likely to have him.
- If it's late-breaking enough, this will tank a large portion of the competition
- Conversely, suppose McDavid will be playing, and a recent AHL call-up appears to be skating on his line during warm-ups — that player is likely low-cost with big potential upside in terms of fantasy points.
- This type of information is exceedingly valuable — though exceedingly time consuming to be on top of.

DFS — Final comments

- In DFS, you're competing against other players.
- Because of high variability in daily performance, and heavy reward for the top of the leaderboard (in larger competitions), it's worth noting that selecting players that nobody else has is one of the core tenets of 'winning big'
- This leads to a bit of a paradox...if everyone is predicting well, everyone will field similar teams, in which case you have higher expected winnings if you select a player predicted to do poorly who happens to do well that night.
- This type of reasoning can drive you crazy...but you can gain some insight by properly investigating the variability in your predictions (think prediction intervals rather than point estimates).

MAIN TAKEAWAY

- If you're interested in Data Science, there are many fun sandboxes out there for you to hone your skills.
- Pursue your non-DS interests using DS! As a statistician, I have zero background in things like web scraping, but I come from a sports-oriented family, and have always participated in fantasy sports.
- From playing around with building models for fantasy sports in my spare time, I've learned countless tricks and techniques that have improved my coding skills.
- These things almost always translate, in some shape or form, into improvements in my core research (finding interesting data sets, improving my open source software, identifying research opportunities, etc...)