

Competitive Machine Learning

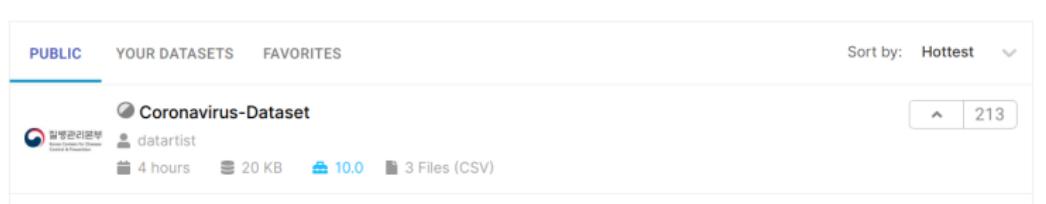
Workshop — March 22, 2021

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?

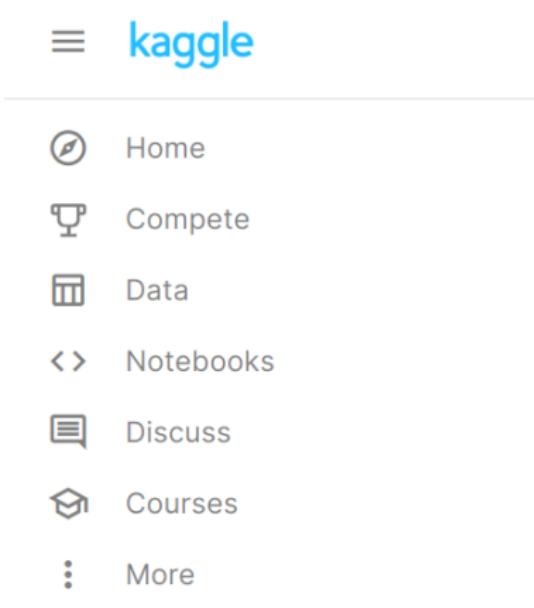
- ▶ Kaggle.com is an online data science competition hosting platform
- ▶ Beyond competitions, it also serves as an excellent repository for interesting data sets.



- ▶ 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



The image shows a screenshot of the Kaggle website's navigation menu. The menu is located on the left side of the page and includes the following items:

- Home
- Compete
- Data
- Notebooks
- Discuss
- Courses
- More

- ▶ First up, compete...

Kaggle — Competitions

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

All Competitions

			All Categories ▾	Default Sort ▾
	Active (Not Entered)	Completed	InClass	
#DFDC	Deepfake Detection Challenge Identify videos with facial or voice manipulations Featured • a month to go • Code Competition • 2019 Teams			\$1,000,000
M5	M5 Forecasting - Accuracy Estimate the unit sales of Walmart retail goods Featured • 4 months to go • 229 Teams			\$50,000
	  Google Cloud & NCAA® ML Competition 2020-NCAAM Apply Machine Learning to NCAA® March Madness® Featured • 14 days to go • 569 Teams			\$25,000
	  Google Cloud & NCAA® ML Competition 2020-NCAAW Apply Machine Learning to NCAA® March Madness® Featured • 15 days to go • 293 Teams			\$25,000

Kaggle — Competitions

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

Your Competitions

Active	Closed	Pinned	Hosted	
M5	M5 Forecasting - Accuracy Estimate the unit sales of Walmart retail goods Featured • 4 months to go • 229 Teams			\$50,000

- ▶ 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

The screenshot shows the 'Overview' page of the 'M5 Forecasting - Accuracy' competition on Kaggle. At the top, it says 'Featured Prediction Competition'. The title 'M5 Forecasting - Accuracy' is displayed with a subtitle 'Estimate the unit sales of Walmart retail goods'. A large image of a city skyline at night is in the background. To the right, it shows '\$50,000 Prize Money'. Below the title, it says 'University of Nicosia · 229 teams · 4 months to go (4 months to go until merger deadline)'. A navigation bar at the bottom includes 'Overview' (which is underlined), Data, Notebooks, Discussion, Leaderboard, Rules, Team, My Submissions, and Submit Predictions.

Overview	
Description	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
Evaluation	
Timeline	
Prizes	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

The screenshot shows the 'M5 Forecasting - Accuracy' competition page on Kaggle. At the top, there's a banner with a city skyline at night and the text '\$50,000 Prize Money'. Below the banner, the competition title 'M5 Forecasting - Accuracy' is displayed, along with the subtitle 'Estimate the unit sales of Walmart retail goods'. It also shows 'University of Nicosia · 229 teams · 4 months to go (4 months to go until merger deadline)'. The navigation bar includes tabs for Overview, Data, Notebooks, Discussion, Leaderboard, Rules, Team, My Submissions, and Submit Predictions. The 'Overview' tab is circled in red. The main content area has two sections: 'Description' and 'Evaluation'. The 'Description' section contains a note about it being one of two complementary competitions and provides a detailed explanation of the challenge. The 'Evaluation' section describes the metrics used for forecasting accuracy.

Featured Prediction Competition

M5 Forecasting - Accuracy

Estimate the unit sales of Walmart retail goods

University of Nicosia · 229 teams · 4 months to go (4 months to go until merger deadline)

Overview Data Notebooks Discussion Leaderboard Rules Team My Submissions Submit Predictions

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

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

The screenshot shows the 'M5 Forecasting - Accuracy' competition page on Kaggle. At the top, it says 'Featured Prediction Competition' and 'M5 Forecasting - Accuracy'. It features a night-time cityscape background. To the right, it displays '\$50,000 Prize Money'. Below the title, it says 'Estimate the unit sales of Walmart retail goods'. It shows the University of Nicosia logo with '229 teams · 4 months to go (4 months to go until merger deadline)'. Navigation tabs include Overview (which is underlined), Data, Notebooks, Discussion, Leaderboard, Rules, Team, My Submissions, and Submit Predictions. The 'Submit Predictions' button is highlighted in blue.

Overview	
Description	This competition uses a Weighted Root Mean Squared Scaled Error (WRMSSE). Extensive details about the metric, scaling, and weighting can be found in the M5 Participants Guide .
Evaluation	Submission File
Timeline	
Prizes	Each row contains an <code>id</code> that is a concatenation of an <code>item_id</code> and a <code>store_id</code> , which is either <code>validation</code> (corresponding to the Public leaderboard), or <code>evaluation</code> (corresponding to the Private leaderboard). You are predicting 28 forecast days (F1-F28) of items sold for each row. For the <code>validation</code> rows, this corresponds to <code>d_1914 - d_1941</code> , and for the <code>evaluation</code> rows, this corresponds to <code>d_1942 - d_1969</code> . (Note: a month before the competition close, the ground truth for the <code>validation</code> rows will be provided.) The files must have a header and should look like the following:

```
id,F1,...,F28  
HOBBIES_1_001_CA_1_validation,0,...,2  
HOBBIES_1_002_CA_1_validation,2,...,11  
...
```

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

The screenshot shows the 'M5 Forecasting - Accuracy' competition page on Kaggle. At the top, it says 'Featured Prediction Competition'. The competition title is 'M5 Forecasting - Accuracy' with the subtitle 'Estimate the unit sales of Walmart retail goods'. A large image of a city skyline at night is displayed. On the right, it shows '\$50,000 Prize Money'. Below the main header, it says 'University of Nicosia · 229 teams · 4 months to go (4 months to go until merger deadline)'. The navigation bar includes 'Overview' (which is underlined), 'Data', 'Notebooks', 'Discussion', 'Leaderboard', 'Rules', 'Team', 'My Submissions', and 'Submit Predictions'. The 'Timeline' section is currently selected. The 'Description' section states: 'June 1, 2020 - Full Training Labels Released. Participants will be provided with the actual values of the 28 days of data used for scoring performance (Public test set). The private test set will remain unchanged, and will still be used to determine the winners.' The 'Evaluation' section states: 'June 23, 2020 - Entry deadline. You must accept the competition rules before this date in order to compete.' The 'Timeline' section states: 'June 23, 2020 - Team Merger deadline. This is the last day participants may join or merge teams.' It also notes: 'June 30, 2020 - Final submission deadline.' The 'Prizes' section states: 'All deadlines are at 11:59 PM UTC on the corresponding day unless otherwise noted. The competition organizers reserve the right to update the contest timeline if they deem it necessary.'

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).

The screenshot shows the 'M5 Forecasting - Accuracy' competition page on Kaggle. The top banner features a night-time cityscape of New York City with the text '\$50,000 Prize Money'. Below the banner, the competition title is 'M5 Forecasting - Accuracy' and the subtitle is 'Estimate the unit sales of Walmart retail goods'. It shows 'University of Nicosia · 229 teams · 4 months to go (4 months to go until merger deadline)'. Navigation links include Overview, Data, Notebooks, Discussion, Leaderboard, Rules, Team, My Submissions, and Submit Predictions. The 'Overview' tab is selected. The 'Prizes' section table lists:

Description	Prize
Description	1st Place - \$25,000
Evaluation	2nd Place - \$10,000
Timeline	3rd Place - \$5,000
Prizes	4th Place - \$3,000
	5th Place - \$2,000

An additional note states: 'An additional \$5,000 will be granted to the highest performing student team on the leaderboard at the end of the competition. A student team is one for which at least 50% of team members are current full-time students. If interested append _STU to your team name.'

Prizes will be distributed during the M5 Conference in December 2020, held in New York City, NY, USA.

Kaggle — Competitions

- Now let's explore the top menu

The screenshot shows the Kaggle interface for the 'M5 Forecasting - Accuracy' competition. At the top, there's a banner for the competition with a city skyline background, showing '\$50,000 Prize Money'. Below the banner, the competition title 'M5 Forecasting - Accuracy' is displayed, along with the subtitle 'Estimate the unit sales of Walmart retail goods'. To the left of the title is the logo of the University of Nicosia, and to the right is a note about 229 teams and 4 months left until merger deadline. The navigation bar includes tabs for Overview (which is underlined), Data, Notebooks, Discussion, Leaderboard, Rules, Team (circled in red), My Submissions, and Submit Predictions. The main content area has a header 'Overview'. Under 'Description', it says: '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](#)'. Under 'Evaluation', it discusses the challenge of predicting weather and business forecasts. Under 'Timeline', it asks how much camping gear will sell each month in a year. Under 'Prizes', it mentions the Makridakis Open Forecasting Center (MOFC) at the University of Nicosia.

Kaggle — Competitions

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

The screenshot shows the 'M5 Forecasting - Accuracy' competition page on Kaggle. The top banner features a night-time cityscape background with the text 'Featured Prediction Competition', 'M5 Forecasting - Accuracy', 'Estimate the unit sales of Walmart retail goods', and '\$50,000 Prize Money'. Below the banner, there's a logo for 'University of Nicosia' and the text '232 teams · 4 months to go (4 months to go until merger deadline)'. A navigation bar includes 'Overview', 'Data' (which is underlined), 'Notebooks', 'Discussion', 'Leaderboard', 'Rules', 'Team', 'My Submissions', and 'Submit Predictions'. The main content area has a section titled 'Data Description' containing text about predicting item sales at stores. Below that is a 'Files' section listing several CSV files: 'calendar.csv', 'sales_train_validation.csv', 'sample_submission.csv', 'sell_prices.csv', and 'sales_train_evaluation.csv'. Each file entry includes a brief description and a link to the 'Evaluation' tab for more info.

Featured Prediction Competition

M5 Forecasting - Accuracy

Estimate the unit sales of Walmart retail goods

\$50,000
Prize Money

University of Nicosia · 232 teams · 4 months to go (4 months to go until merger deadline)

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.

The screenshot shows the 'M5 Forecasting - Accuracy' competition page. At the top, it says 'Featured Prediction Competition'. The main title is 'M5 Forecasting - Accuracy' with the subtitle 'Estimate the unit sales of Walmart retail goods'. A large image of a city skyline at night is displayed. To the right, it shows '\$50,000 Prize Money'. Below the title, it says 'University of Nicosia · 232 teams · 4 months to go (4 months to go until merger deadline)'. The navigation bar includes 'Overview', 'Data', 'Notebooks' (which is underlined), 'Discussion', 'Leaderboard', 'Rules', and 'Team'. A 'New Notebook' button is located on the right side of the bar.

The screenshot shows a list of notebooks from the 'M5 Forecasting - Accuracy' competition. The top navigation bar has tabs for 'Public', 'Your Work', 'Shared With You', and 'Favorites', with 'Public' selected. It also includes 'Sort by Hotness' and a search bar 'Search notebooks'. The notebook list includes:

Rank	User	Notebook Title	Outputs	Languages	Tags	Actions
120		Back to (predict) the future - Interactive M5 EDA	13h ago	time series, beginner, eda, data visualization		Rmd 24
1		M5_Forecasting_eda	43m ago			Py
13		M5 Competition : EDA + Models	6h ago	0.86729		Py
5		Ensemble Starter	17h ago	0.84849		Py 4

Kaggle — Competitions

- ▶ Discussion is a forum area

The screenshot shows the 'Discussion' tab of the Kaggle M5 Forecasting Accuracy competition page. The top banner features a city skyline at night and the text '\$50,000 Prize Money'. Below the banner, the competition title 'M5 Forecasting - Accuracy' and subtitle 'Estimate the unit sales of Walmart retail goods' are displayed. A logo for 'University of Nicosia' indicates 232 teams have joined. The navigation bar includes 'Overview', 'Data', 'Notebooks', 'Discussion' (which is underlined), 'Leaderboard', 'Rules', and 'Team'. On the right, there are buttons for 'My Submissions' and 'New Topic'. The main content area displays a list of 34 topics, each with a user profile picture, topic title, poster's name, posting time, last comment by, last comment time, and a reply count. The topics listed are:

Topic ID	User	Topic Title	Poster	Posted	Last Comment by	Last Comment Time	Replies
5	Addison Howard	Welcome to the M5 Competitions!		2 days ago	Igor Krasovskiy	19h ago	5
3	Addison Howard	New to Machine Learning or Kaggle?		2 days ago		2d ago	3
4	Addison Howard	External Data/Pre-Trained Models Disclosure Thread		2 days ago	Addison Howard	2d ago	0
1	Addison Howard	Looking for a Team Megathread		2 days ago	Justin Black	2d ago	10

Kaggle — Competitions

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

The screenshot shows the 'M5 Forecasting - Accuracy' competition page on Kaggle. The top banner features a night cityscape background with the text 'Featured Prediction Competition', 'M5 Forecasting - Accuracy', 'Estimate the unit sales of Walmart retail goods', '\$50,000 Prize Money', and the University of Nicosia logo with '232 teams · 4 months to go (4 months to go until merger deadline)'. Below the banner, there are navigation tabs: Overview, Data, Notebooks, Discussion, Leaderboard, Rules, Team, My Submissions, and Submit Predictions. The Leaderboard tab is active. The main content area displays the Public Leaderboard. It includes a note about the leaderboard being calculated with 50% test data and final results based on 50% test data. There are buttons for 'Raw Data' and 'Refresh'. A legend indicates four categories: 'In the money' (green), 'Gold' (orange), 'Silver' (grey), and 'Bronze' (brown). The leader table lists five entries:

#	Team Name	Notebook	Team Members	Score	Entries	Last
1	Ethan			0.82948	13	1h
2	SVJ24			0.83404	7	1h
3	Arsa Nikzad			0.83736	10	3h
4	satoshi			0.83746	4	2d
5	Sepid K			0.83779	2	3h

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 withholding 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

PLAYER	FANDUEL				STATS						
	EP	COST	VALUE	SHOTS	GOALS	AST	PPG	PPA	±L	BLK	MIN
Patrice Bergeron C BOS @ CLB 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...)