Capstone 2: Milestone Report

Problem Statement:

The second capstone will be the creation of a working daily fantasy model forecast system.

This is an advance of the first capstone, which was a seasonal model, based on historical data, whereas this is a daily forecast with rolling updates. This adds multiple layers of complexity as it requires both the regular daily database updates as well as designing new features and using models for forecasting instead of a regression or other method.

Both Machine Learning techniques and Financial Analyst methods will be attempted in order to provide more depth to the model beyond basic statistics recorded by the National Hockey League.

Dataset:

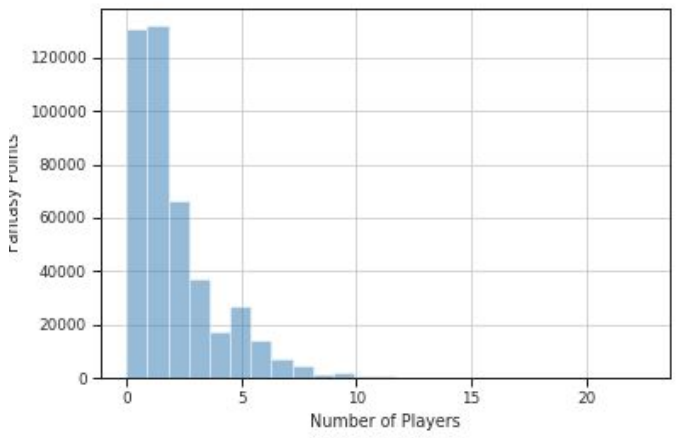
The initial data was compiled using a well known api scraper, and compiler documented here. The flat file may also be downloaded from the website evolving-hockey, but this method was chosen to further develop more advanced api scraping and r scripting skills. Given that that was not the focus, here the work has been done on the compiled csv files that were saved from the api and R, with only the work done in python shown.

The primary data source for that api is the official, if albeit completely undocumented and unpublished, api of National Hockey League. The files cover 10 seasons of data, but the quality is significantly lesser before the 2015 season.

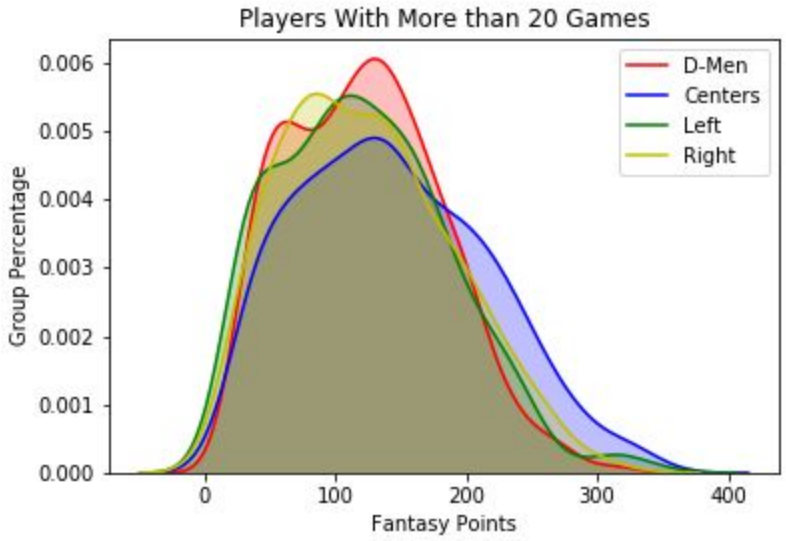
A large challenge was the transforming the API data into model ready form, and creating summary statistics from that. Again, this was assisted by the evolving wild process, as well as some pandas functions to create the moving averages.

Initial Findings:

While most of the initial exploratory stats were congruent with the first capstone, a level of difficulty is added by the strong leftward tail in the target dependant variable, fantasy points as seen in the figure below.

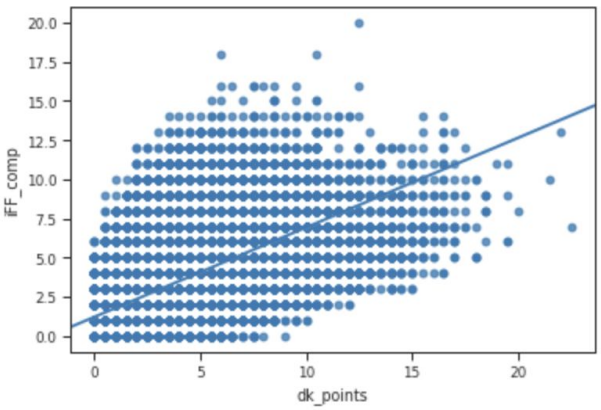


This makes sense given the designers of daily fantasy hockey want to create a challenge in selecting winning combinations, but is radically different than the season long product where there is a much more centered distribution. For contrast here is the season long version, as reported in the first capstone.

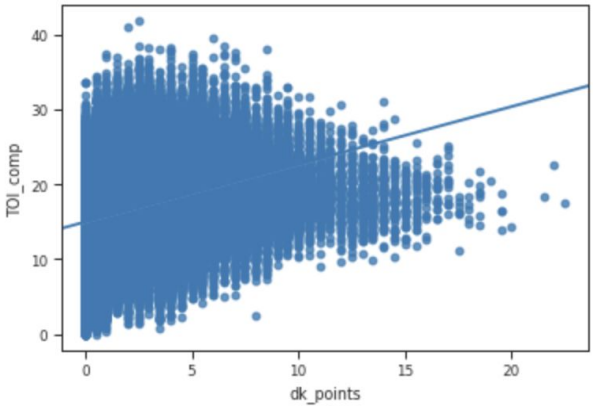


That said, the same variables as before show strong correlation to the dependant variable, namely Time on Ice and Fenwick, a value that measures the various shot attempts. Below are the respective plots for those variables.

Fenwick



Time On Ice



Initial Model: Finance:

At the heart of the project is the idea that you can forecast future performance on past results. Perhaps nowhere else has this been as intensely explored as the financial sector. Worth noting is that these financial methods have also been heavily critiqued, perhaps most notably by the legendary Burton Malkiel, but where they fail in financial markets they could very well succeed in this use case. The most frequent critique of 'Chart watchers' and 'market timers' is that they are using prior information to fill in future assumptions, while the efficient market theory states that there is far more current information that is readily available and therefore the market price at current time isn't based on past pricing. In this case, we do not have more current information or insight into why a player is seeing more time; they might have been moved due to injury or a trade. Even if we do have some of that information for some players via say Twitter or trade trackers, coverage isn't global and we don't have an efficient way to find out if a player will see an increase in ice time. Therefore, finding a moving average might well be an effective signal to show that a player's points will soon increase as well.

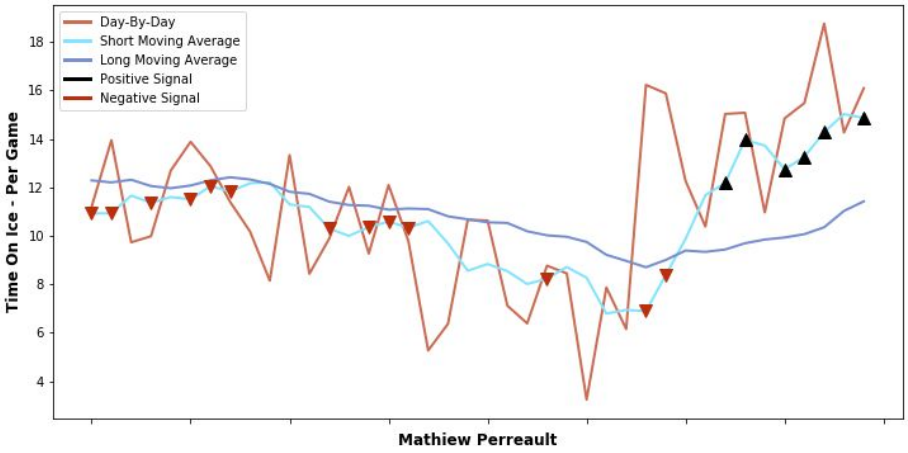
Here our hypothesis is that if the average time, power play time, or possession metrics have increased across the past five games over the previous 20, the player will continue to both maintain that higher standing and, due to the nature of those metrics, also see an increase in over all points.

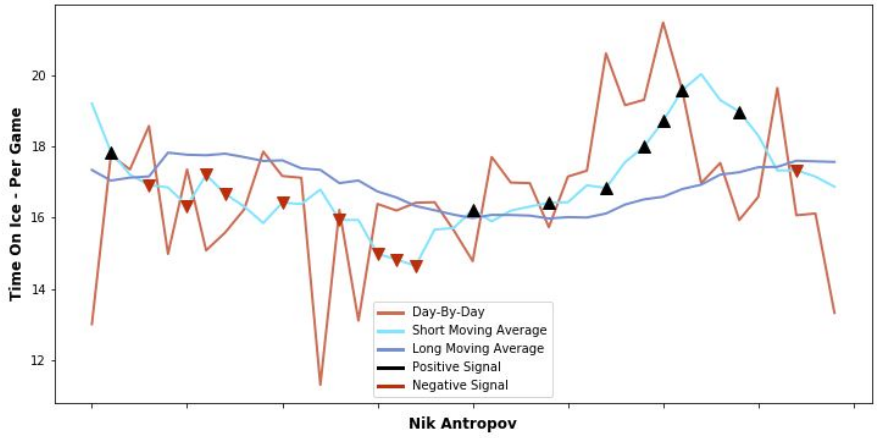
Market theorists have proposed the idea of a 'dual moving average crossover' or a 'moving average convergence'; this is a lagging indicator that shows upward momentum, where if the shorter average is above the value of the larger moving average, this is a positive trend signal and a time to 'buy' or select such a player.

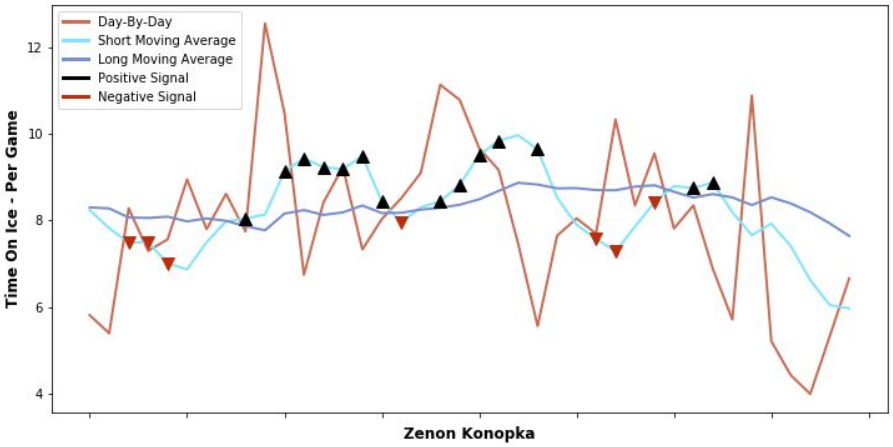
In order to limit noise, these metrics were created for time on ice, power play time on ice, Fenwick, and over all draftkings points, and then back tested before being added to the over all package. Credit for the base code and the visual here go to this excellent DataCamp tutorial, 'Python For Finance: Algorithmic Trading'.

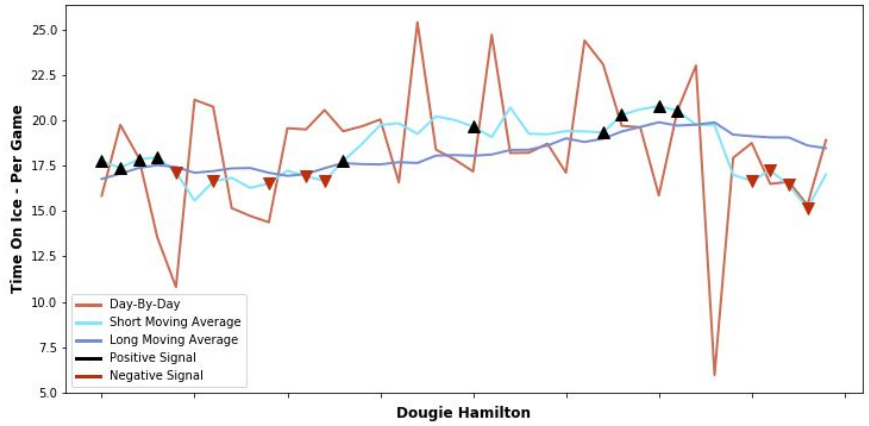
A snapshot of this first application to time on ice looked quite promising. Below is the charting of four randomly selected players with the signal markers overlayed, with black arrows representing a signal of an upcoming positive trend and red signalling an upcoming negative trend.

Moving Averages and Signal Indicators - Random Sample of Players and Games




Time On Ice - Per Game
20
18
16
14
12
Day-By-Day
Short Moving Average
Long Moving Average
Positive Signal
Negative Signal
Nik Antropov


Time On Ice - Per Game
12
10
8
6
4
Day-By-Day
Short Moving Average
Long Moving Average
Positive Signal
Negative Signal
Zenon Konopka


Time On Ice - Per Game
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
Day-By-Day
Short Moving Average
Long Moving Average
Positive Signal
Negative Signal
Dougie Hamilton

This seemed promising, and a grid search was generated to find the best performing windows, which ended up being over 3 and 20 games respectively, and then back tested on the past games. Unfortunately, while this further analysis showed that while these signals are effective, the overall yield was approximately 3 seconds in time, less than 1/10th to 1/100th of the average shift. When utilized in an OLS model with total points as the dependant variable, this was slightly more effective, finding the coefficient at 8 seconds of time, almost 1/4th a shift, but the long moving average was a much better indicator, being worth more than 30 seconds: a complete additional shift.

OLS Regression Results

Dep. Variable:	TOI_comp	R-squared:	0.972
Model:	OLS	Adj. R-squared:	0.972
Method:	Least Squares	F-statistic:	3.092e+06
Date:	Tue, 17 Sep 2019	Prob (F-statistic):	0.00
Time:	18:59:51	Log-Likelihood:	-8.7465e+05
No. Observations:	353268	AIC:	1.749e+06
Df Residuals:	353264	BIC:	1.749e+06
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
short_mavg	0.4604	0.004	108.273	0.000	0.452	0.469
long_mavg	0.5331	0.004	135.472	0.000	0.525	0.541
positions	-0.1409	0.012	-11.415	0.000	-0.165	-0.117
toi_signal	0.1363	0.016	8.594	0.000	0.105	0.167

Omnibus:	46756.872	Durbin-Watson:	2.003
Prob(Omnibus):	0.000	Jarque-Bera (JB):	578353.067
Skew:	0.140	Prob(JB):	0.00
Kurtosis:	9.262	Cond. No.	89.6

This was predictably less successful in the case of points, showing less than .04 points increase for the signal, and .06 points for the long average.

OLS Regression Results

Dep. Variable:	dk_points	R-squared:	0.541
Model:	OLS	Adj. R-squared:	0.541
Method:	Least Squares	F-statistic:	1.041e+05
Date:	Tue, 17 Sep 2019	Prob (F-statistic):	0.00
Time:	19:07:55	Log-Likelihood:	-7.2697e+05
No. Observations:	353268	AIC:	1.454e+06
Df Residuals:	353264	BIC:	1.454e+06
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
short_mavg	0.0545	0.003	19.454	0.000	0.049	0.060
long_mavg	0.0647	0.003	24.993	0.000	0.060	0.070
positions	-0.0280	0.008	-3.451	0.001	-0.044	-0.012
toi_signal	0.0442	0.010	4.229	0.000	0.024	0.065

Omnibus:	128101.103	Durbin-Watson:	2.003
Prob(Omnibus):	0.000	Jarque-Bera (JB):	522647.738
Skew:	1.779	Prob(JB):	0.00
Kurtosis:	7.781	Cond. No.	89.6

When the same methods were applied to Fenwick the results were far more encouraging, with significant increases both in Fenwick events and overall fantasy points, as indicated in two OLS models.

Again the long moving average is most successful, but the signal is in fact effective, showing a .4 point bump per game.

What's perhaps most interesting here is dependant variable here actually decreases with the short moving average in both instances, but the long average and the Fenwick Signal seems to be worth almost a half a point, which in terms of an 8 man fantasy hockey team where difference between cashing in a tournament or not is often far less than that.

OLS Regression Results

Dep. Variable:	iFF_comp	R-squared:	0.692
Model:	OLS	Adj. R-squared:	0.692
Method:	Least Squares	F-statistic:	1.986e+05
Date:	Tue, 17 Sep 2019	Prob (F-statistic):	0.00
Time:	19:40:19	Log-Likelihood:	-6.8397e+05
No. Observations:	353268	AIC:	1.368e+06
Df Residuals:	353264	BIC:	1.368e+06
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
short_mavg	-0.0188	0.005	-3.976	0.000	-0.028	-0.010
long_mavg	0.9484	0.004	237.691	0.000	0.941	0.956
positions	-0.0969	0.006	-15.028	0.000	-0.110	-0.084
FF_signal	0.2628	0.009	30.540	0.000	0.246	0.280

Omnibus:	35945.756	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	55779.263
Skew:	0.757	Prob(JB):	0.00
Kurtosis:	4.223	Cond. No.	13.4

OLS Regression Results

Dep. Variable:	dk_points	R-squared:	0.564
Model:	OLS	Adj. R-squared:	0.564
Method:	Least Squares	F-statistic:	1.141e+05
Date:	Tue, 17 Sep 2019	Prob (F-statistic):	0.00
Time:	19:40:33	Log-Likelihood:	-7.1802e+05
No. Observations:	353268	AIC:	1.436e+06
Df Residuals:	353264	BIC:	1.436e+06
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
short_mavg	-0.0891	0.005	-17.141	0.000	-0.099	-0.079
long_mavg	0.8332	0.004	189.642	0.000	0.825	0.842
positions	-0.1410	0.007	-19.850	0.000	-0.155	-0.127
FF_signal	0.4067	0.009	42.926	0.000	0.388	0.425

Omnibus:	105376.837	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	368300.817
Skew:	1.493	Prob(JB):	0.00
Kurtosis:	7.014	Cond. No.	13.4

Next Steps:

For the final leg we'll rejoin the tables and create a composite model for daily use, and test it using historical tracking and team generators to see how it would perform in real time.