IEOR4523 Data Analytics - Final Project Report
Tofu Chill: Zhirui Cheng zc2503, Ziwei Li zl2853, Zhe Xi zx2286, Yuchen Fei yf2515

Outline:

- Overview
- Innovation
- Data description
- Key steps and Analysis (Main Line)
- Text mining experiment
- Findings
- Limitations

Overview:

In our project, we would like to find a way to see what makes contributions to an NBA star's paycheck and if we are able to make predictions on it. We collect data on players' salaries and performance statistics over the years with their biological statistics using machine learning algorithms to see if NBA stars' performances are the dominant determinants in how they are paid. We also try to use twitter as a text data source to help with the prediction to see if it captures any non-objective factors. We come to a conclusion that performance has a large positive correlation with salary but with performance alone we are unable to make an accurate prediction on NBA players' salaries.

Keywords: predictive analysis, salary decision, NBA, machine learning, text mining

Innovations:

1. **Extract data from multiple sources with various methods:** We scrap player salary data from Hoopshype website, download player performance history from NBA official website and player bio stats data from kaggle. Plus, we use twitter API to extract 1 month of text about NBA stars related. We construct a rather resourceful and comprehensive master dataset despite of all the hard work in data cleaning.
2. **Invent better quality variable as prediction label:** Instead of using the absolute salary with or without adjustment of CPI like most of previous articles have done, we use salary as a percentage of salary cap (the total salary limit that a team can spend on its players in a given season) in corresponding year as the prediction label. We construct a more time-stationary dependent variable to in machine learning algorithm. This brings better predictive power in a time-series data and also help normalize salary data across the year.
3. **Develop different assumptions of salary decision step by steps:** To better approach the NBA players' salaries, we start with baseline assumption of salary based on past one year performance and continue with adding the past year salary to make it a 'time series model'. We then develop 'time-lag model' to capture more previous performance to capture potential long-term effect.

4. **Apply multiple machine learning methods and grid search to achieve the best results:** We implement decision tree, random forest, bagging, support vector regression, xgboost, gradient boosting regressor and neutral network methods beside linear regression methods in predictions. For each of the machine learning methods we tested with grid search to select the best parameters for each model, and used cross validation RMSE (Root Mean Squared Error) as a comparing line among different models.
5. **Apply text-mining to capture non-objective decision making:** With an assumption that the player salary decision would also be affected by subjective issues like its reputation and public comment, we use twitter to capture the sentiment and emotions towards player to help predict their salaries.

Data Description:

Data Sources:

Data theme	Data Source	Time Range	Approach
Player Performance	NBA official website	Season 1996 - 2018	Download
Player Salary	Hoopshype	Season 1990 - 2019	Web-Scraping
Player Bio	Kaggle Dataset	Static: Heights, weights, ...	Download
Comments on players	Twitter	Raw tweet text	Twitter API

Basic Data Information: The final merged dataset has **8117** rows and **34** columns, with features are:

- 5 Basic Features:'Year', 'Name', 'TEAM', 'tenure', 'Salary_Weight'.
- 26 Performance Features:'GP', 'W', 'MIN', 'PTS', 'FGM', 'FGA', 'FG%', '3PM', '3PA', '3P%', 'FTM', 'FTA', 'FT%', 'OREB', 'DREB','‘REB', 'AST', 'TOV', 'STL', 'BLK', 'PF', 'FP', 'DD2', 'TD3', '+/-', 'Position' (see stats glossary in appendix).
- 3 Bio Features: 'Height', 'Weight','AGE'.

Key Steps and Analysis (Main Line):

- **Extract and gather data**
  -refer notebook

- **Merge and clean data**
  -refer notebook

- **Machine Learning Algorithms under different models**

- **Models and Behind Assumptions:**

Model	Assumption	Adding Features
Baseline Model	The rationale of salary decision does not change over time	N/A.
Auto-regression Model	Last year's salary might also serve as a reference of decision	Salary_l
Time-Lagged feature Model	Previous years' performance and salary both contribute to the new decision	PTS_l1, 3PM_l1, W_l1, ...

- **CV RMSE results across models and algorithms:**

Algorithm	Baseline	Auto-Regression	TimeLag1	TimeLag2
Linear Regression	0.05822	0.04309	0.04303	0.04356
Decision Tree	0.06238	0.04633	0.04647	0.04775
Random Forest	0.05382	0.04165	0.04163	0.04312
Bagging	0.05451	0.04278	0.04816	0.04339
Gradient Boost Machine	0.05388	0.04201	0.04281	0.04337
XGBoost	0.05391	0.04222	0.04245	0.04319

- **Final Model**: Based on model simplicity and cv RMSE, our final model is **Random Forest with Auto-Regression Assumption.**
  - Testing Data Performance: RMSE = 0.04221

The result is not optimistic after timing the scale factor salary cap, which is usually around $ 90,000,000. That means our mean error for each player is around $3,600,000, which is really high. We analyzed this bad result in the last section of the report.

Text Mining Experiment:

- **Use Twitter API** to access the text data about players
  - refer notebook

- **Clean data**

- refer notebook

- **Sentiment Analysis**
    - refer notebook

- **Bagging method including sentiment attributes**
    - Instead of using our original best model Random Forest under Auto-Regression Assumption, we decided to apply bagging method under Auto-Regression Assumption to the the new data set with sentiment features due to a limit in the data size.
    - We compared Root Mean Squared Error (RMSE) of original and text model, and the results are as follows:

Model	Original Model	Text model
Bagging	0.066073565	0.065481707

We are happy to see that there is an increase in model performance when we include the text features. However, since this is a very limited size data, on the one hand we may foresee a bigger improvement from text features given a larger size but on the other hand we are not sure if this input would be significant. We would expect later research could focus more on looking for a better source of text data to extend the available time frame.

Findings & Reflections:

There may be a long lasting stereotype that high performance means bigger paychecks. Actually, this “counterintuitive” phenomenon shows us the real payment condition in the NBA, which exactly the fact our analytics results want to illustrate.

- Indeed, we can look into a lot of real scenarios, for instance:
    - Stephen Curry won the MVP in 2014-2015 season, but he earned only 10 million dollars, at that time, the top salary was 20 million.
    - Pascal Siakam only earns 2.3 million right now, but he can get 25.1 PTS and 8.6 rebounds per game.

- To understand this conclusion deeply, we can think about the following aspects:
    - Many players will get injured during the long-term contract, which greatly influence their performance in the future.
    - The Rookie Contracts are relatively low, but the players can have awesome performance.
    - Teams sometimes sign a player by judging his potential, but the expectation will fail to achieve.

To sum up, our analytic proves that the salaries of NBA players are not the accuracy indicator for their performance as the time-lag effect, accidents, contract rules and so forth. Considering the fact mentioned above, we can revise analytics further. To improve our models, we can divide our sample with more details, such as set up a contrast group that eliminates the players engage in the Rookie Contract, to see whether the relation between performance and salary grow more accurately and objectively.

Limitations:

- Text Mining:
  - Limited time-span of text available
  - Limited Player mentioned available
  - fan's opinion might matter less than the professional sports critics
- Peak Year effect and free agent:
  - Free player compromise on a small package to wait for a peak year contract
- Business endorsement
  - Advertisement may also be a factor that could help explain if a player is popular and has public attention

Appendix:

Exhibit 1: the cleaned text dataset we get from twitter API

date	Tweets	User	User_statuses_count	user_followers	User_location	User_verified	fav_count	rt_count	tweet_date
2019-11-15	Mans gonna drain the 3 in your eye and	xavi 🤴🟫	13089	980	Naguabo, Puerto Rico	FALSE	0	3344	2019-11-15 23:59:15
2019-11-15	nes Harden or Chris P. Bacon 🥓? https	realist guy you know 🅱️	20751	487	West Mifflin, PA	FALSE	0	9	2019-11-15 23:59:08
2019-11-15	Mans gonna drain the 3 in your eye and	St. Andrew✝🕊	10726	411		FALSE	0	3344	2019-11-15 23:58:49
2019-11-15	Mans gonna drain the 3 in your eye and	Shawn	1659	385		FALSE	0	3344	2019-11-15 23:58:35
2019-11-15	out James Harden right now are just po	Marcelo Rugiero	63339	742	Ciudad Autónoma de Buenos Aire	FALSE	0	25	2019-11-15 23:58:33
2019-11-15	en you can behold the glory of a future a	Patrick Araya	15779	493	Philly	FALSE	0	148	2019-11-15 23:58:21
2019-11-15	out James Harden right now are just po	jugg	173067	840	Nawf	FALSE	0	25	2019-11-15 23:57:19
2019-11-15	Rockets five-game winning streak! 👀 4	- Issael Guillen 😊	159741	1440		FALSE	0	4	2019-11-15 23:54:57
2019-11-15	awhi Leonard, leads a reverse alley oop	WORLDWIDE CLYDE ‼️	83521	1644	504	FALSE	0	1098	2019-11-15 23:54:24
2019-11-15	un between Kawhi and James Harden 👀	🧘‍♂	15325	747		FALSE	0	3883	2019-11-15 23:54:21
2019-11-15	Mans gonna drain the 3 in your eye and	n8🌊	4019	433	301 • Towson U	FALSE	0	3344	2019-11-15 23:52:48
2019-11-15	nes Harden or Chris P. Bacon 🥓? https	🌒'd ⛄	57842	1998	Atlanta, Georgia	FALSE	0	9	2019-11-15 23:52:43
2019-11-15	Mans gonna drain the 3 in your eye and	Yzaya🅱️	114	28	Tucson, AZ	FALSE	0	3344	2019-11-15 23:52:15
2019-11-15	etball players of all-time comfortably an	—	9288	1638		FALSE	9	1	2019-11-15 23:51:26
2019-11-15	rden ROTY: Coby White DPOY: Anthony	.	53516	12918		FALSE	0	1	2019-11-15 23:51:15
2019-11-15	White DPOY: Anthony Davis 6MOY: Lou	P r o p h e ☨	18001	694	North Carolina, USA	FALSE	5	1	2019-11-15 23:50:54
2019-11-15	Mans gonna drain the 3 in your eye and	🤘🔥	3663	77	Pasadena, TX	FALSE	0	3344	2019-11-15 23:50:25
2019-11-15	Mans gonna drain the 3 in your eye and	Agozie Anyamene	16761	390		FALSE	0	3344	2019-11-15 23:50:02
2019-11-15	41.6 points (in 38.2 minutes/gm) 45.7%	Melanated Bearded Ken Do	63389	672	IG:johanbravo1 👻:johnnyrose_1	FALSE	0	257	2019-11-15 23:49:36
2019-11-15	Mans gonna drain the 3 in your eye and	Lorenzo	53550	527	In the gym	FALSE	0	3344	2019-11-15 23:49:30
2019-11-15	Mans gonna drain the 3 in your eye and	haydo	8580	489	HB, Cali	FALSE	0	3344	2019-11-15 23:48:05
2019-11-15	Mans gonna drain the 3 in your eye and	Ry Dolla $ign ⚡	47619	1803	New Haven, CT	FALSE	0	3344	2019-11-15 23:47:48
2019-11-15	Mans gonna drain the 3 in your eye and	LordJonny	16167	310		FALSE	0	3344	2019-11-15 23:47:16
2019-11-15	Mans gonna drain the 3 in your eye and	luc	41482	966	lynn,MA	FALSE	0	3344	2019-11-15 23:46:56
2019-11-15	on is right now. Better ball handler, way	playoffchef	774	38		FALSE	1	0	2019-11-15 23:46:44
2019-11-15	You commented about stats saying is "1	.	53516	12918		FALSE	0	1	2019-11-15 23:45:36
2019-11-15	Mans gonna drain the 3 in your eye and	Parm	3651	387		FALSE	0	3344	2019-11-15 23:44:44
2019-11-15	Mans gonna drain the 3 in your eye and	fabiandavalos52	247	93		FALSE	0	3344	2019-11-15 23:44:31
2019-11-15	em looking like james harden at the aw	Caleb👑	1886	603	nicki followed 8/7/19	FALSE	5	0	2019-11-15 23:44:14
2019-11-15	ames ➖➖➖ 41.6 PPG 8.8 APG 6.8 RI	Naz 😜	67275	556		FALSE	0	510	2019-11-15 23:42:14
2019-11-15	s Harden with the thermal radiation fit.	Dan Favale	38030	8859	IG: danfavale	TRUE	0	3	2019-11-15 23:40:44
2019-11-15	Mans gonna drain the 3 in your eye and	Coziest of Cozies, Trick Da	50598	1715	San Mo, HOME4BREAKINHEARTSBABY	FALSE	0	3344	2019-11-15 23:40:43
2019-11-15	Conference Finals with a series lead: htt	Paully T.	79261	589	Las Vegas, NV	FALSE	0	2	2019-11-15 23:40:12
2019-11-15	out James Harden right now are just po	JustinCase Dykstra	760	35		FALSE	0	25	2019-11-15 23:39:45
2019-11-15	Mans gonna drain the 3 in your eye and	😴	30194	302	Toronto, Ontario	FALSE	0	3344	2019-11-15 23:39:44
2019-11-15	earing what appears to be an aluminum	Chiaki	6358	33	Grenoble, France	FALSE	0	5	2019-11-15 23:39:12
2019-11-15	Mans gonna drain the 3 in your eye and	Day Beezy🦅	7472	149		FALSE	0	3344	2019-11-15 23:38:55
2019-11-15	Wonder_Kid4 Not everything is about Jar	GE🏐RGE	17940	881	Texas	FALSE	0	1	2019-11-15 23:38:04
2019-11-15	earing what appears to be an aluminum	GOOCH MAYNE KRIK!!!	191254	2011	KRIKLAHOMA	FALSE	0	5	2019-11-15 23:38:04

Exhibit2 : Data cleaning code and the file results

- James Harden
- Anthony Davis
- LeBron James
- Damian Lillard
- Giannis Antetokounmpo

```
import nltk
from nltk.corpus import PlaintextCorpusReader
tweets_root = excel_code
player_lst=[]
for i,player in enumerate(players):
    player_file = "{}.*".format(player)
    player_data = PlaintextCorpusReader(tweets_root,player_file)
    player_lst.append([player,player_data.raw()])
dff=comparative_emotion_analyzer(player_lst,object_name='player',print_output=False)
```

Exhibit 3: Sentiment and Emotion score for Players

player	Negative	Positive	Surprise	Anticipation	Trust	Joy	Disgust	Anger	Sadness	Fear
Dennis Smith Jr.	0.01461814	0.0391453	0.01168153	0.02358615	0.02862787	0.02317243	0.0020686	0.00567041	0.00335844	0.00507822
Isaiah Thomas	0.01380506	0.03895986	0.01112727	0.02311967	0.02780787	0.02299196	0.00204336	0.00541327	0.00349761	0.00449046
Austin Rivers	0.01362465	0.03768157	0.01083496	0.02262887	0.02548913	0.02220958	0.00233304	0.00577449	0.00364901	0.00484044
Andre Drummond	0.01755269	0.03906142	0.0092389	0.01970721	0.02284365	0.01908236	0.00233711	0.00552224	0.00369231	0.0045525
Otto Porter Jr.	0.01724495	0.0386561	0.00928574	0.01982094	0.0228675	0.01909074	0.00234882	0.00553331	0.00376054	0.00462056
Rodney Hood	0.01428876	0.03506177	0.00958851	0.02013378	0.02316283	0.0196742	0.00338836	0.00660542	0.00439108	0.00578235
Dario Saric	0.01446834	0.03453422	0.00959818	0.02022892	0.02312175	0.01967711	0.00346555	0.006651	0.00444794	0.00591944
Paul Millsap	0.0141091	0.03534375	0.00967755	0.02044942	0.02353605	0.02006516	0.00304068	0.00635703	0.00414335	0.00553003
Reggie Jackson	0.01503829	0.03441757	0.00922209	0.02018944	0.02232568	0.01910923	0.00209593	0.00761387	0.00347037	0.00439742
Jaylen Brown	0.01252684	0.02633633	0.00640477	0.01700343	0.01208383	0.01313441	0.00337959	0.01139188	0.00640899	0.01089401
Zach Randolph	0.01230772	0.02623732	0.00634392	0.01708889	0.01210498	0.01313555	0.00327755	0.01108708	0.00614119	0.01082522
Jusuf Nurkic	0.01205615	0.02667743	0.00708567	0.01758171	0.01267851	0.01370456	0.00319591	0.01097963	0.00604279	0.01060537
Dion Waiters	0.01276538	0.02567949	0.00653992	0.01715507	0.01236168	0.01316908	0.00337407	0.01116758	0.00585151	0.01124832
Bojan Bogdanovic	0.01368219	0.02431143	0.0083738	0.01619516	0.01105707	0.01247762	0.00305295	0.01276838	0.00521701	0.01239455
Avery Bradley	0.01220938	0.02438953	0.00686464	0.01610519	0.01115296	0.01243904	0.00299806	0.01030532	0.00482696	0.0098293
Jeff Teague	0.01182819	0.02490101	0.00642364	0.01656866	0.01161101	0.01292664	0.00308235	0.01027866	0.00478641	0.01018678
Taurean Prince	0.01183541	0.02489027	0.00641467	0.01632484	0.01135097	0.01286693	0.00308623	0.01022339	0.00479012	0.0098517
Enes Kanter	0.01364741	0.02639601	0.00591854	0.01756483	0.01261717	0.01290547	0.00400646	0.01080261	0.00500278	0.01056943
Hassan Whiteside	0.01315048	0.02670285	0.00832367	0.01787786	0.01232599	0.01456331	0.00291266	0.00759446	0.00419705	0.01070186
Jonathon Simmons	0.0130224	0.02656735	0.00814107	0.0177171	0.01236298	0.01437855	0.002957	0.00764755	0.00425509	0.01055063
Rondae Hollis-Jefferson	0.01177858	0.02542023	0.00671454	0.01636748	0.01253715	0.01311649	0.00302175	0.00776069	0.00420961	0.00916946
Jayson Tatum	0.0144847	0.0241201	0.00706175	0.01261542	0.01406028	0.01420025	0.00497573	0.01065583	0.00811378	0.01128344
Steven Adams	0.01719479	0.02683999	0.01129724	0.0150313	0.01277276	0.0167286	0.00558525	0.00972214	0.00697478	0.01341094
Jordan Clarkson	0.0138413	0.02548693	0.00720018	0.01335437	0.01267809	0.01338593	0.00560866	0.00992786	0.00701082	0.0103697
Clint Capela	0.01407721	0.02069614	0.00629737	0.01069885	0.01105381	0.01163009	0.00402564	0.00777984	0.00511139	0.00753346
Marcus Morris Sr.	0.01401317	0.02052657	0.0061305	0.01052965	0.01089589	0.01145775	0.00401625	0.00773284	0.0050484	0.0074748
Rudy Gobert	0.01481717	0.02408172	0.00656457	0.0123789	0.01313786	0.01360458	0.00443599	0.00856229	0.00560497	0.00830931
Buddy Hield	0.01415109	0.02252814	0.00661487	0.01169567	0.01224579	0.01309557	0.0040387	0.0134981	0.00826524	0.01085036

Exhibit 4: Performance Stats Glossary

GP Games Played

W Wins

L Losses

MIN Minutes Played

FGM Field Goals Made

FGA Field Goals Attempted

FG% Field Goal Percentage

3PM 3 Point Field Goals Made

3PA 3 Point Field Goals Attempted

3P% 3 Point Field Goals Percentage

FTM Free Throws Made

FTA Free Throws Attempted

FT% Free Throw Percentage

OREB Offensive Rebounds

DREB Defensive Rebounds

REB Rebounds

AST Assists

TOV Turnovers
STL Steals
BLK Blocks
PF Personal Fouls
FP Fantasy Points
DD2 Double doubles
TD3 Triple doubles
PTS Points
+/- Plus Minus