NBA Match Outcome Prediction

By:
Neil Gibeaut
Kevin Thomas Mathew
Nishant Barma

NBA

- The 2018 NBA finals had about 17.7 million viewers
- $1.4 billion bet on basketball (college+professional) last year
- Deal with MGM may make NBA betting a much larger force in the future
- Lots of data available on NBA games
- **Goal:** Predict the winner of a match **Given:** All match data up until that game.
- Challenging problem - old data may do more harm than good
- Only a fixed number of matchups from which to draw inference

Related Work

- Torres, Renato Amorim. "Prediction of NBA games based on Machine Learning Methods." *University of Wisconsin, Madison*(2013).
    - The goal of this paper is to survey several machine learning methods on a limited set of features. The main contribution was a good feature set starting point for predicting NBA seasons 2006 - 2012. It was determined that linear classifiers are particularly effective at predicting the outcome of an NBA match.

- Hoffman, Lori, and Maria Joseph. "A Multivariate Statistical Analysis of the NBA."
    - Focused on exploring different basketball team and player features. Focuses more on statistical analysis rather than machine learning.

Related Work

- Miljković, Dragan, et al. "The use of data mining for basketball matches outcomes prediction." *Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on*. IEEE, 2010.
    - Uses data mining techniques such as multivariate linear regression to predict the outcome of the games. Besides predicting the actual outcome, the spread for each game was also calculated.

- Beckler, Matthew, Hongfei Wang, and Michael Papamichael. "NBA oracle." *Zuletzt besucht am* 17.20082009.9 (2013).
    - This paper also evaluates several machine learning models. The paper also focuses on player-centric features along with team-centric features to predict the outcome of a basketball game. It also focuses on unsupervised learning methods to predict optimal player positioning. Best accuracy achieved was 73%.

Approach

- Data from stats.nba.com.
- Used nba_py web scraper
- Used existing features to create more complex features
- All features were mean normalized
- We perform feature selection
- The prediction uses data from past games to calculate if the home team wins or loses

Approach

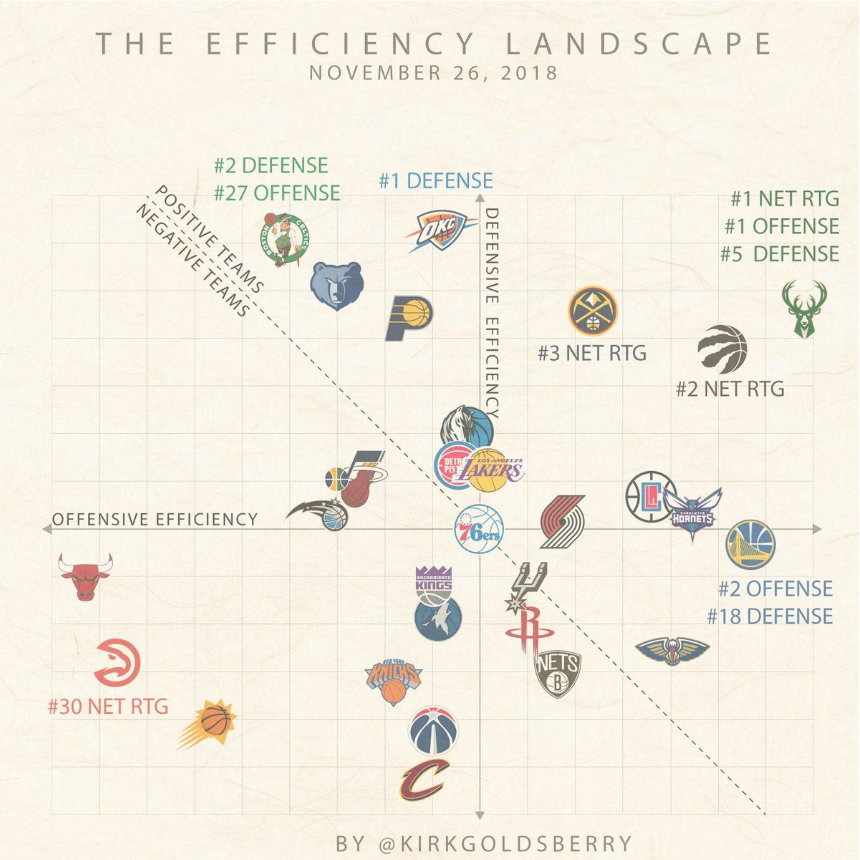
- Prediction models used:
  - Logistic Regression
  - Support Vector Machines
  - Random Forest Classifier
  - AdaBoost Classifier
  - Neural Network
- Evaluation using accuracy on binary prediction
- We compare our results to a baseline of experts - average accuracy 68-71%

Features

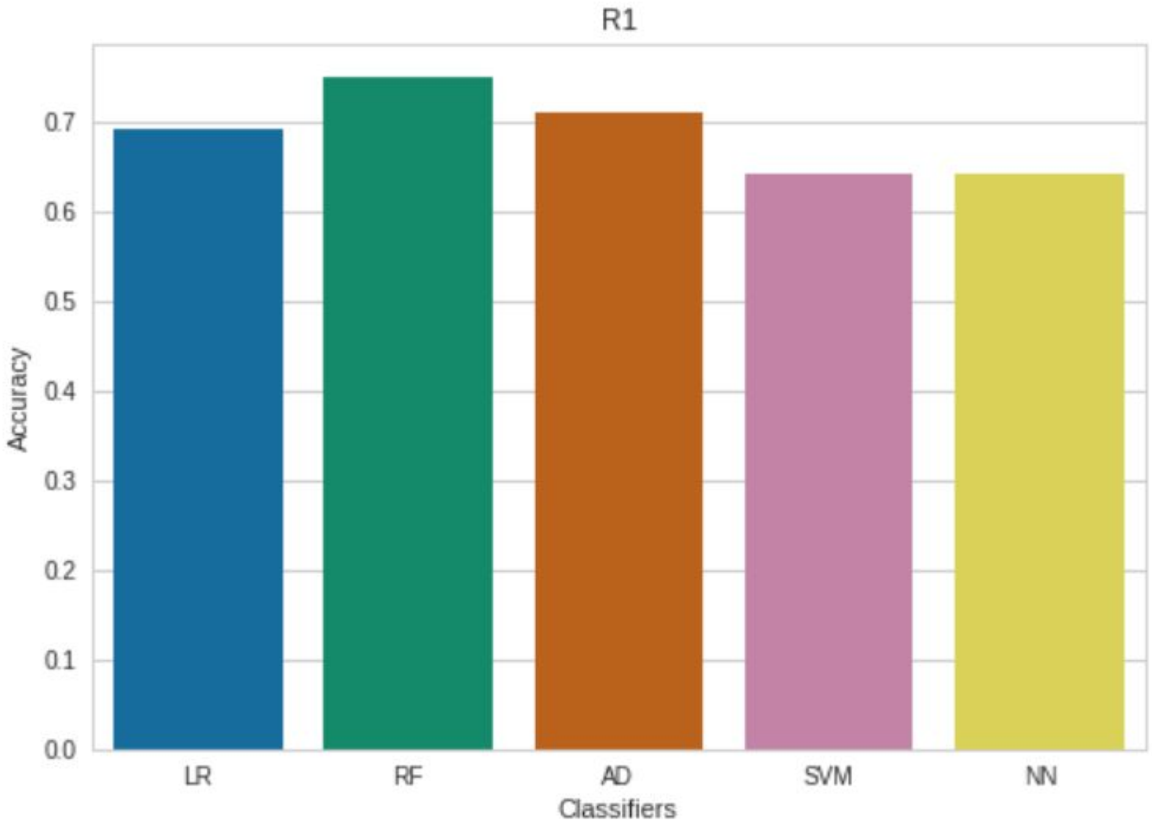
- WIN_PERCENTAGE_A (Last 15 games)
- WIN_PERCENTAGE_B   (Last 15 games)
- DAYS_SINCE_LAST_GAME_A
- DAYS_SINCE_LAST_GAME_B
- EXPERT_RANK_HOME (Aggregate preseason ranks from espn.com, SI.com, etc.)
- EXPERT_RANK_AWAY
- One Hot Encoding of Team Names
- RELEVANT_NET_RTG_HOME
- RELEVANT_NET_RTG_AWAY

Relevant Net Rating

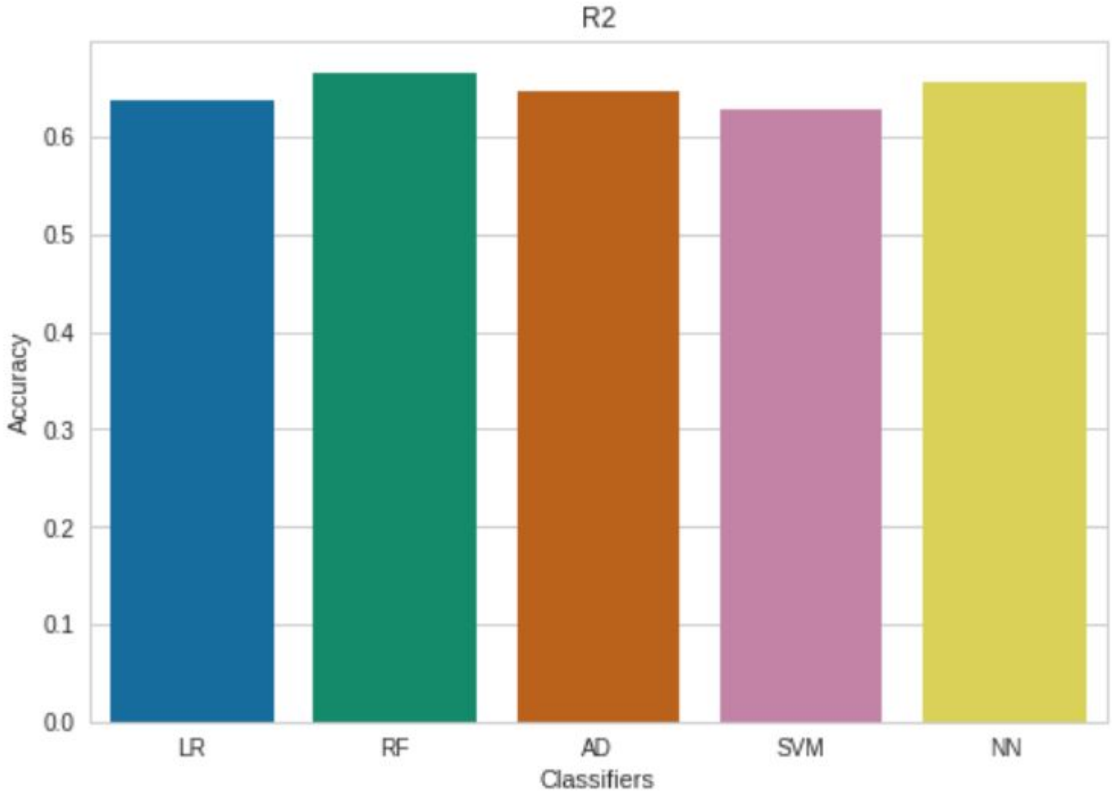
- Net Rating = Points scored per 100 possessions - Points allowed per 100 possessions
- Relevance:
  - Last N Games (N = 5 in our models)
  - Last N Common Opponents (N = 1 in our models)
    - Net rating from past game Team A and Team B have both played against the same opponent



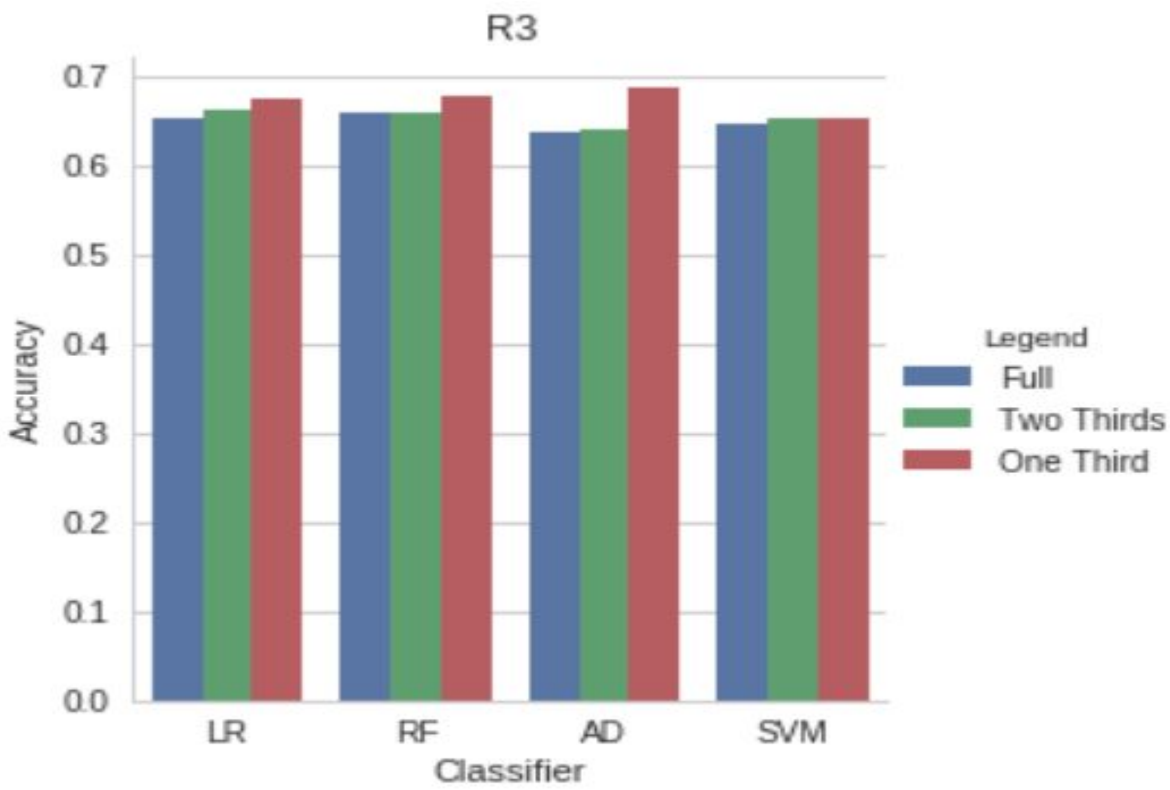
Results - Cross Validation



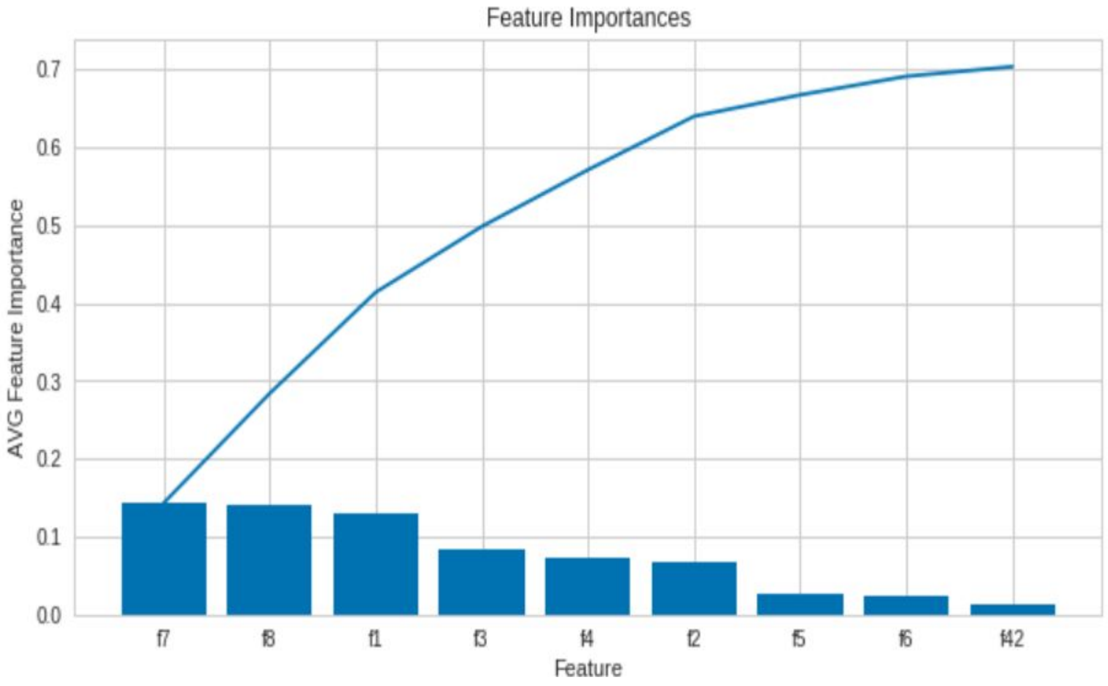
Results - Predicting Far into Future



Results - One Day at a Time



Feature Importance (Random Forest)



F7 - Expert Rank Home

F8 - Expert Rank Away

F1 - Net Rating Home

F3 - Win Percentage Team A

F4 - Win Percentage Team B

F2 - Expert Rank Away

F5 - Days since last game Team A

F6 - Days since last game Team B

Analysis

- Logistic Regression, AdaBoosts, and Random Forest perform best in most cases
- Feature selection improves the accuracy for LogR
- Deciding the “best” features to compute is very empirical
- Significantly better performance on R1:
  - Future data is seen by model
  - Improves ability to infer performance when trained over very “relevant” tuples
- R3 performs better on later slices of the season:
  - Better data available for capturing current season trends
  - Corroborated when looking at the accuracy/day across a season
- More complex representations of current features can improve performance

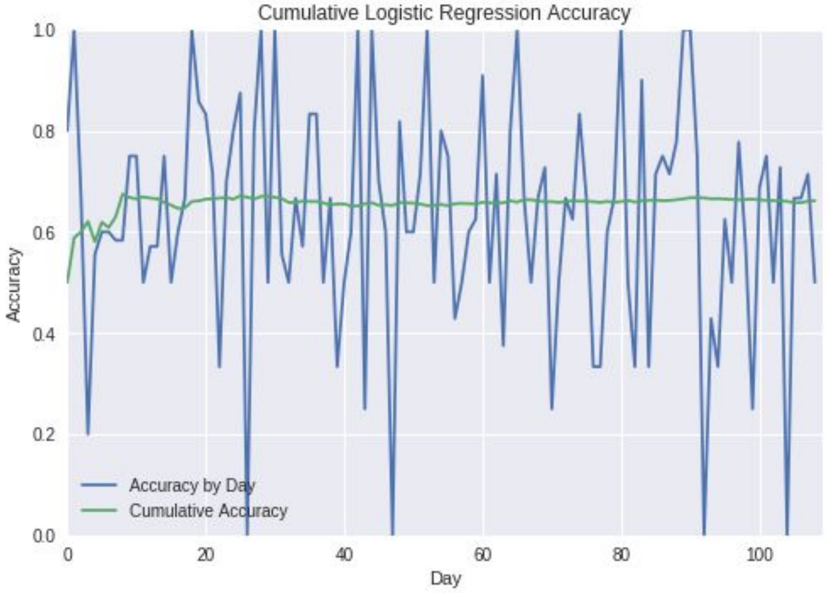
Conclusion

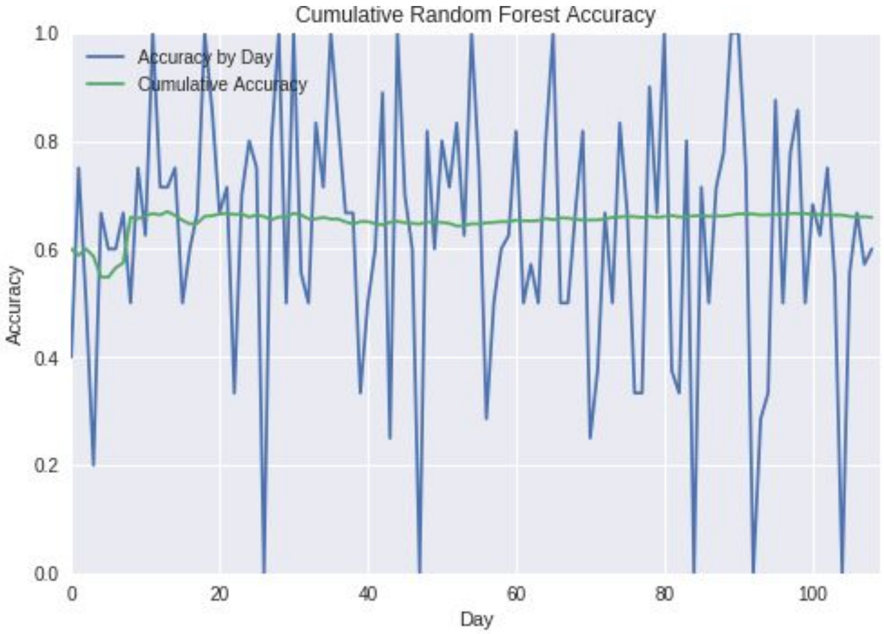
- It is difficult to create a model that outperforms experts, but possible to perform models that perform close to as well
- Having the most recent data is very important for performance
- The model goes out of date quickly
- The models struggle to predict the first games of a season (cold start)

Further Work

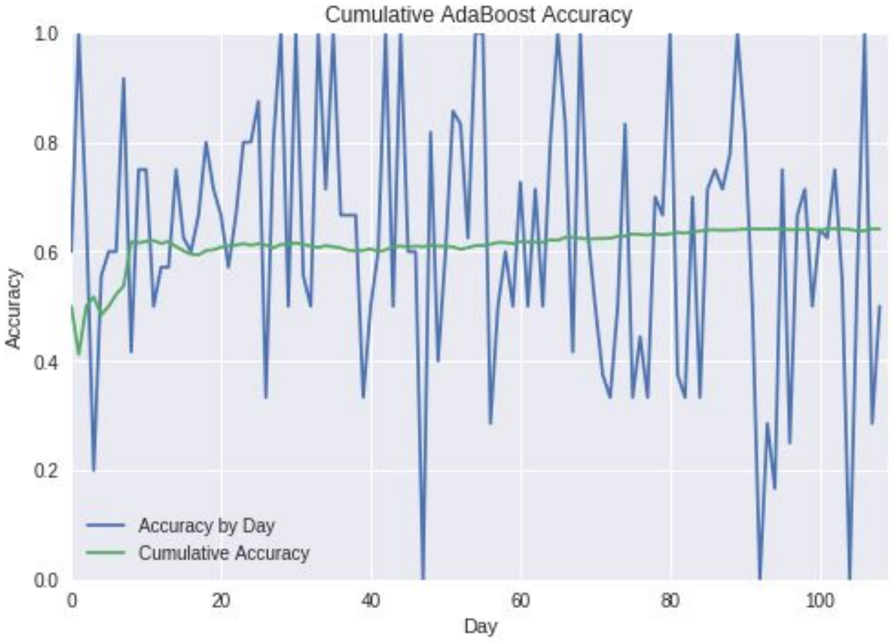
- More empirical evaluation of different feature mixtures and hyperparameters for our features
- Updating expert rankings throughout season
  - Currently only using a preseason ranking
- Expanding predictions to more than just outcome (point spreads, player stats, etc.)

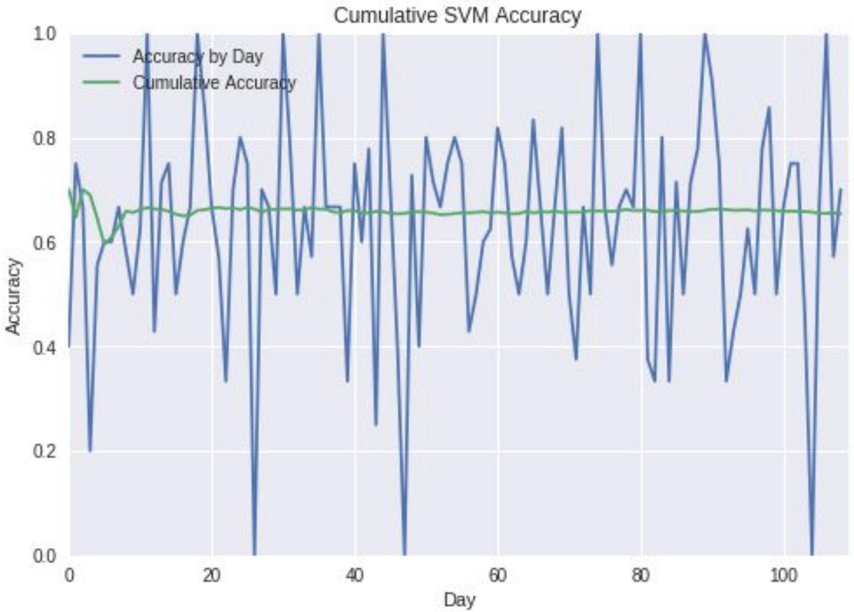
Extras - R3(two thirds)



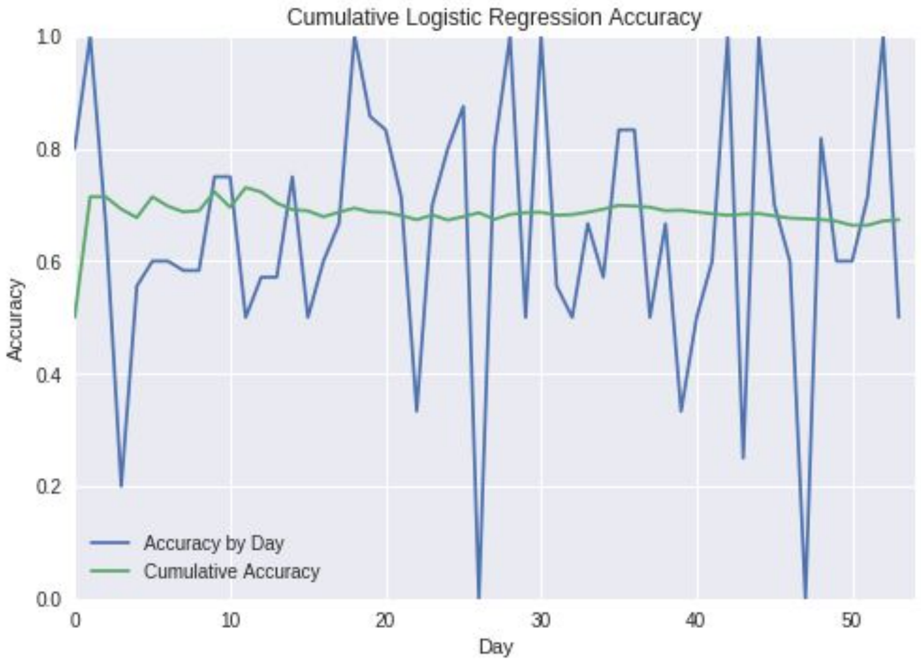


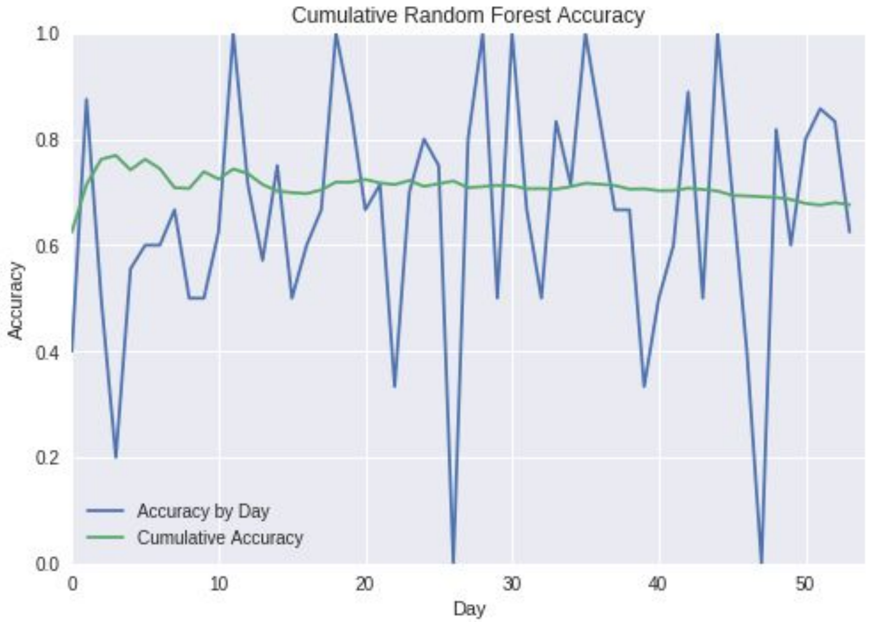
Extras - R3(two thirds)





Extras - R3(one third)





Extras - R3(one third)