The Wisdom of Crowds: A Natural Language Processing Approach to Forecasting Sports Betting Markets Using Social Media Fan Sentiment

Peter Chen, ORFE '19
Advised by Professor René Carmona

Background

As part of the larger movement of using "alternative data" to inform decisions, we implement a proof of concept for **wisdom of crowds**, or the idea that the collective knowledge of a group of people can be used as an alternative to expert opinion. In this study, we use fan sentiment in the form of Reddit submissions and comments to predict wagers on NFL games. The domains are selected for the following reasons:

NFL

- NFL is the most profitable sports league in the world, and NFL betting is the largest sports betting market in the United States, with $3 billion wagered (legally) annually.
- American football is the most popular spectator sport in the United States by rating, allowing a wealth of fan speculation.
- NFL teams play only one game per week, usually on Sunday, giving a defined sampling period to gather fan sentiments leading up to each game.

Reddit

- Reddit is one of the top 5 most visited websites in the United States, a convenient overlap with NFL viewers.
- The upvoting system identifies the most important or agreeable opinions or news headlines, which in turn may contain more predictive information.
- The subreddit system provides separate communities for fans of each team which serve as conveniently predetermined sampling domains.

We focus on two most popular forms of sports betting on the per game level, wagering which team will **win the point spread (WTS)**, a handicap for the team bookkeepers expect will win the game, and whether the combined score will be above or below the **over-under line**, a prediction for the total score set by bookkeepers. Sports betting typically follows the "11 for 10" rule, meaning that a player needs to bet 11 dollars to win 10 dollars, with the bookkeeper keeping the remaining 1 dollar. A prediction model would need to achieve accuracy higher than threshold P to be profitable:

$$10\,P = 11\,(1 - P)$$
$$P = 52.4\%$$

Goal

Consistently obtain a **prediction accuracy** above 52.4% for the binary classification task of selecting the winner of the spread and over-under wagers for NFL games using optimized sources of public sentiment data available on the social news aggregation website Reddit.

Data

1862 total games played from 2012 to 2018 seasons played among 32 teams. For each game, we scrape submissions and comments from each team's subreddit and represent as 32 combinations of feature matrices:

Content

- **Submission titles** from a team's fan community, or subreddit, typically consisting of objective news (team performance, trades, signings, injuries etc.)
- **Comments** posted in response to submissions or other comments, tending to consist of subjective personal opinions

Date Range

- Top 100 top voted **3-days** leading up to gameday
- Top 100 top voted **7-days** leading up to gameday
- Top 100 top voted **14-days** leading up to gameday
- Top 100 top voted **30-days** leading up to gameday

Representation Model

- **Bag of Words** - representing text as frequency of its "stemmed" tokens (removing numbers/symbols/stop words)
- **Term Frequency-Inverse Document Frequency (TF-IDF)** - representing text as frequency of its tokens scaled proportionally to the number of times it appears in a document offset by the number of documents in the corpus containing the token
- **Vader** - representing text as negativity, neutrality, positivity, and composite scores
- **Afinn** - representing text via a single number quantifying its polarity

For each feature matrix, we apply a principal component analysis as a method of dimensionality reduction, projecting onto number of principal components that capture 90% of explained variance. We are predicting for two **target variables**, winner of the spread and over-under.

Approach

For the 32 representations of fan sentiment, we train and test the following 6 machine learning models, each belonging to a distinct learning paradigm.

Generative Models

- Naïve Bayes (Gaussian and Multinomial)
- K-Nearest Neighbors (k as a tunable hyperparameter)

Discriminative Models

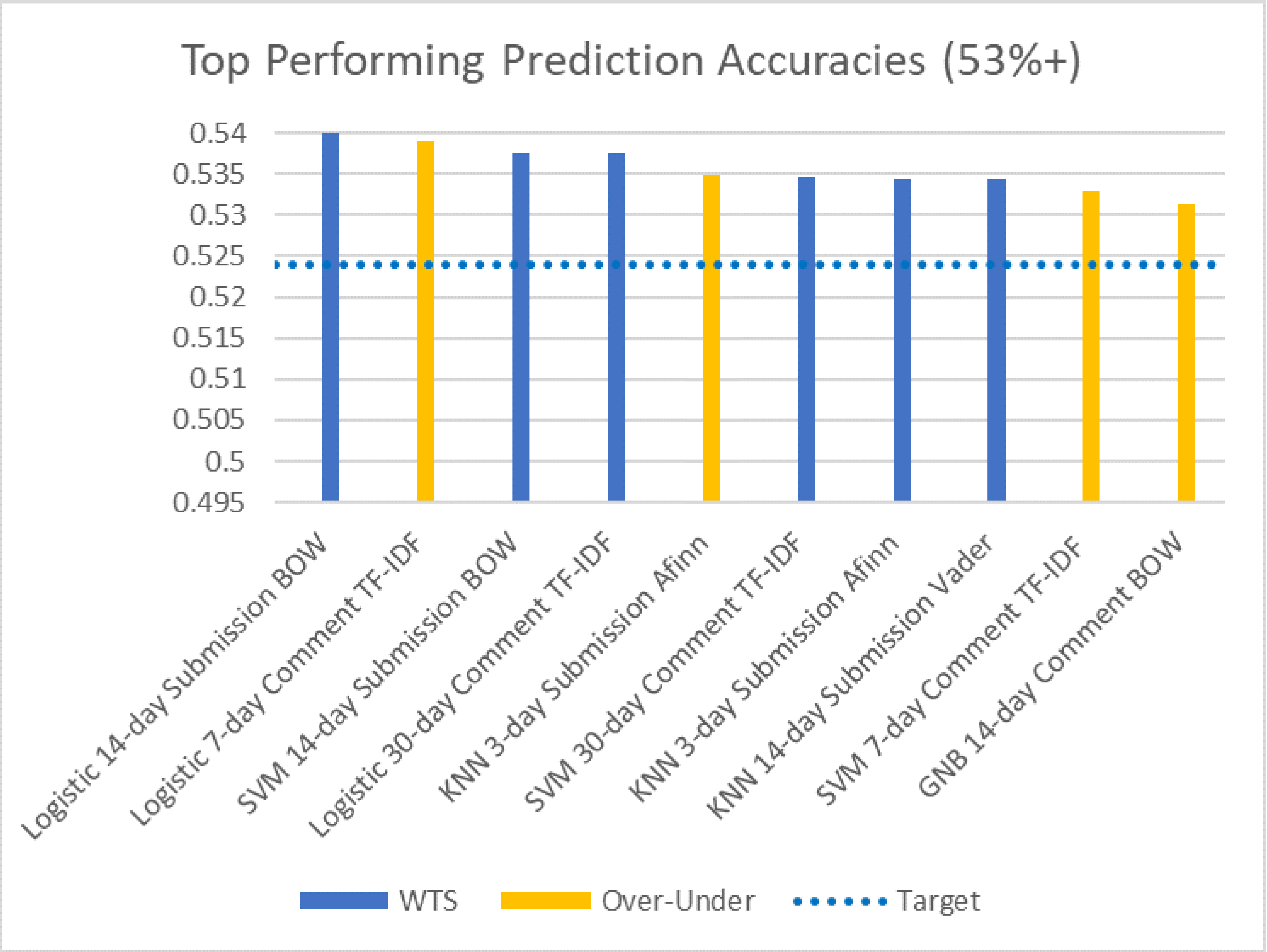
- Logistic Regression (C=1/λ parametrized with L2 regularization)
- Linear Support Vector Machine (C-parametrized)

Decision Machines

- Random Forest (depth of tree as tunable hyperparameter)
- LSTM Neural Networks (ordered by day of submission)

Results

Each feature matrix and model parameter is tested **100** times (for a total of 500 times) to arrive at an average prediction accuracy. We attain similar levels of success predicting both winners of the spread and the over-under, reaching the high 53's and well above the threshold for profitability.



We see that discriminative models, logistic regression and linear support vector machines, coupled with bag of words and TF-IDF representations were able to generally outperform all other models. K-nearest neighbor models work exceptionally well with low dimensional scoring models that capture basic sentiments from text.

Average Prediction Accuracies

3-day	0.5077	Submissions	0.5083
7-day	0.5050	Comments	0.5069
14-day	0.5095	naïve Bayes	0.5025
30-day	0.5100	KNN	0.5120
BOW	0.5084	Logistic	0.5118
TF-IDF	0.5112	Linear SVM	0.5105
Vader	0.5064	Random Forest	0.5034
Afinn	0.5045	LSTM*	0.5090

* Due to time constraints, LSTM networks were only tested 5 times, but bag of words representations of submissions titles were able to reach accuracies above 53%.

Future Work

Future work with the problem as presently framed involves improving **data signals** (i.e. with in-game performance statistics), incorporating **time-sensitive approaches**, and **fine-tuning existing models.** More interesting implications of this study involve the application of this general framework to other fields of research, such as equity markets or political strategy.

CENTER FOR STATISTICS AND MACHINE LEARNING
PRINCETON UNIVERSITY