College Football Bowl Game Predictor

Evan Cheshire, Austin Childs, Thomas Leung

PROBLEM

The goal of this project is to predict the outcomes of bowl games at the end of the college football season. The games are usually against two evenly matched teams, and the result are gambled on. A game, the College Football Bowl Pick'em is played by many fans. Thus, a predictor to pick the right games can win someone a little money and provide that person with some bragging rights.

MODELS

We worked with two models for the majority of the time:
1. Score Regression – Linear Regression based off game points
2. Differential Regression – Linear Regression based off score differential

Other models, such as SVM and logistic classifiers weren't as strong. Testing data would be one year of bowl games, and training data would be all other years outside of that test year.
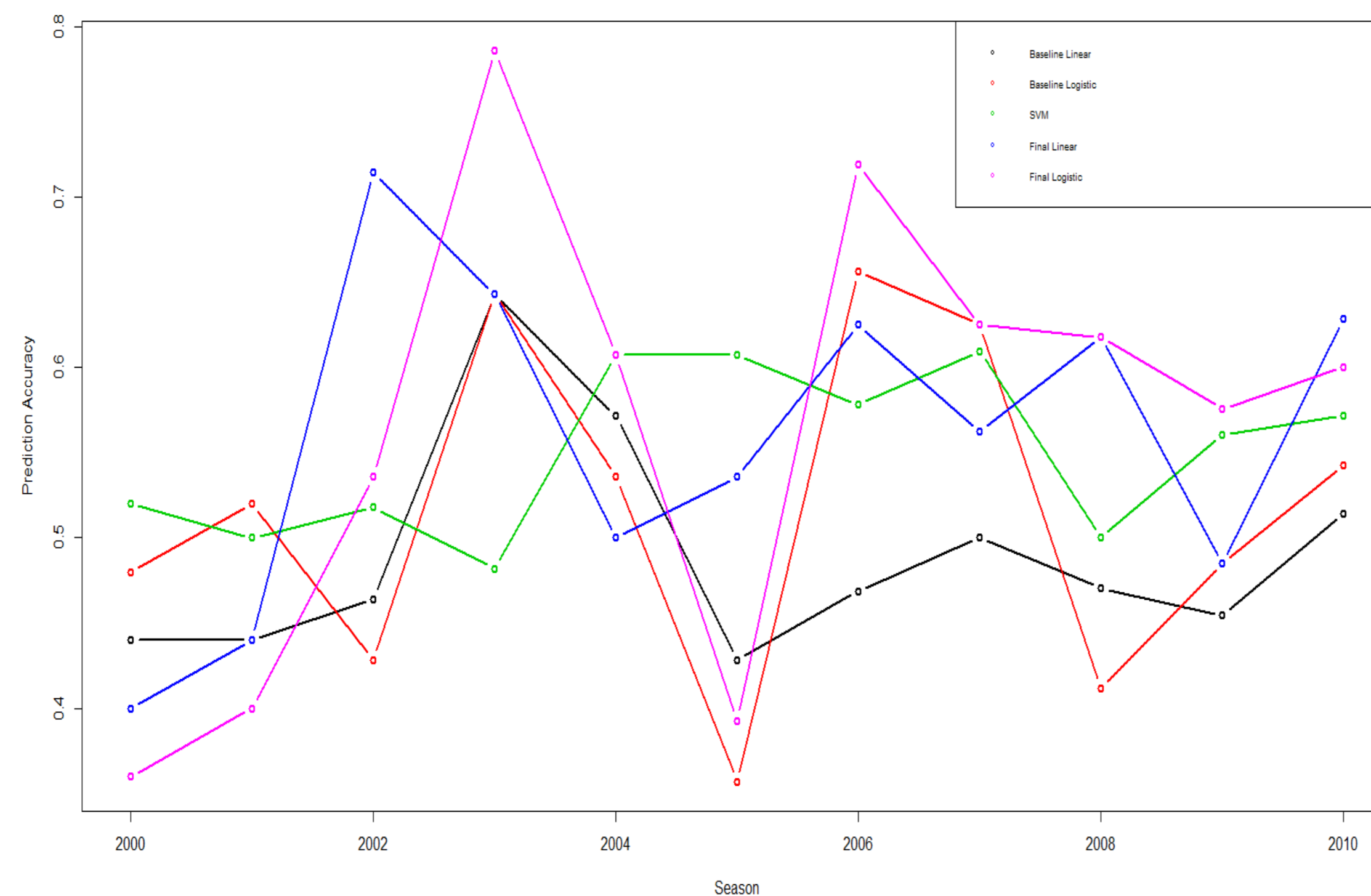
DATA & METHODS

Our data contains regular season and bowl game statistics from 2000 to 2010. We created a MySQL database and used Python and MATLAB to query and process this data. We averaged the regular season game statistics for each team, which we used as features for our models. To test our model, we ran a cross-validation on a season by season basis. Thus our features were selected to maximize our season by season model.
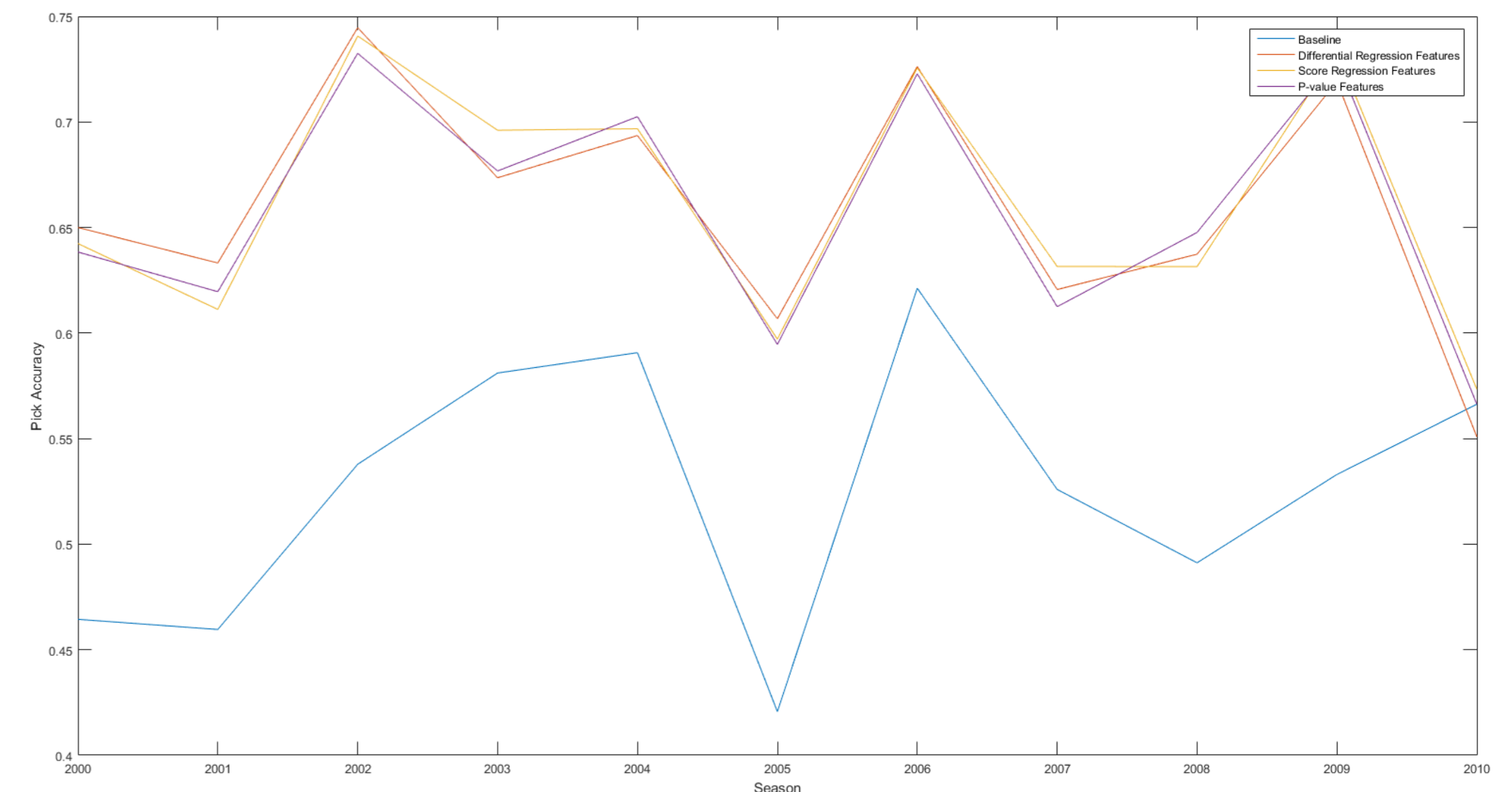
INITIAL FEATURES

Average Points
Average Penalty Yards
Average Yards Per Carry
Average Yards Per Pass
Completion Percentage
Average Touchdowns
Average Field Goals
Field Goal Percentage
Average Kickoff Return
Average Punt Return
Total Turnovers
Average First Downs
Points Variance
Penalty Yards Variance
Rush Yards Variance
Pass Yards Variance
Touchdowns Variance
First Downs Variance

OTHER MODEL

DIFFERENTIAL MODEL

RESULTS & CONCLUSIONS

We found that the differential regression yielded better results than the score regression model. The linear regression on the differentials had the highest prediction accuracy of 66%. For comparison, we looked at the results of Microsoft's Cortana software, accurate to 67%, and Nate Silver's Elo model, accurate to 70%.

REFERENCES

[1] Hamadani, Babak. Predicting the outcome of nfl games using machine learning. URL http://cs229.stanford.edu/proj2006/BabakHamadani-PredictingNFLGames.pdf.
[2] Qin, Dennis. Predictions and Rankings in College Football. URL http://cs229.stanford.edu/proj2009/Qin.pdf
[3] Hamann, John. What It Takes To Win: A Machine Learning Analysis of the College Football Box Score. URL http://cs229.stanford.edu/proj2011/Hamann-WhatItTakesToWin.pdf

[4] http://thenationalchampionshipissue.blogspot.com/2005/08/
[5] http://www.cortanapredictions.com/
[6] http://projects.fivethirtyeight.com/2015-nfl-predictions/
Simulation Packages:
1. Matlab
2. R
3. MySQL
4. Python