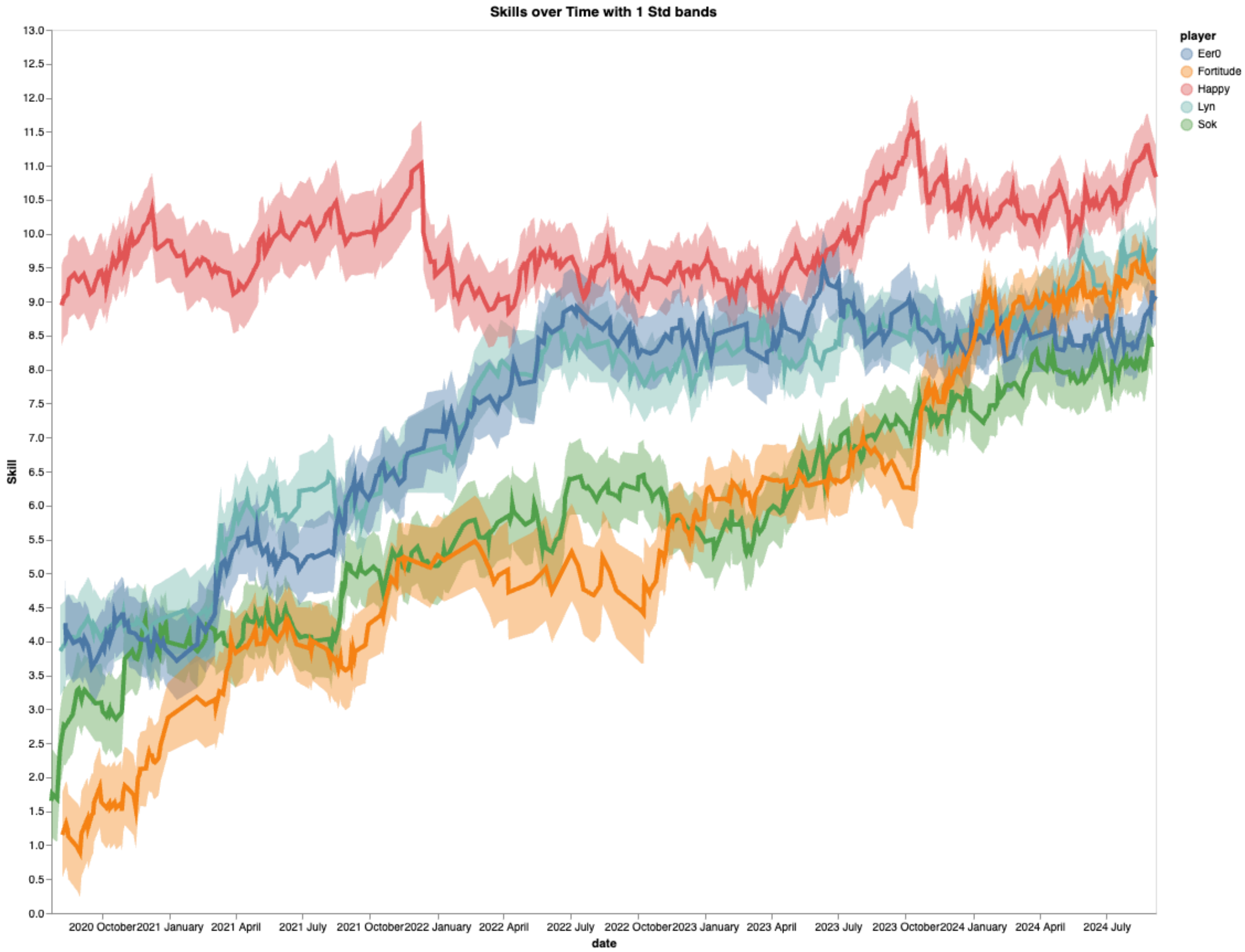
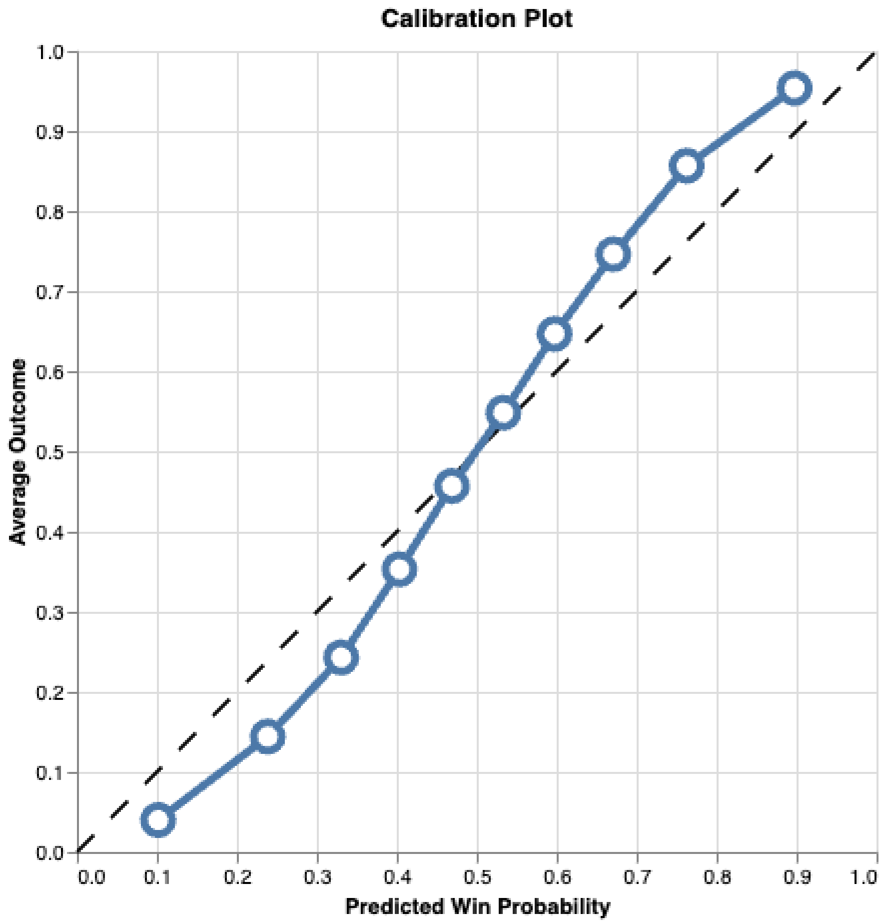
Apply TrueSkillThroughTime algo on 85777 games:

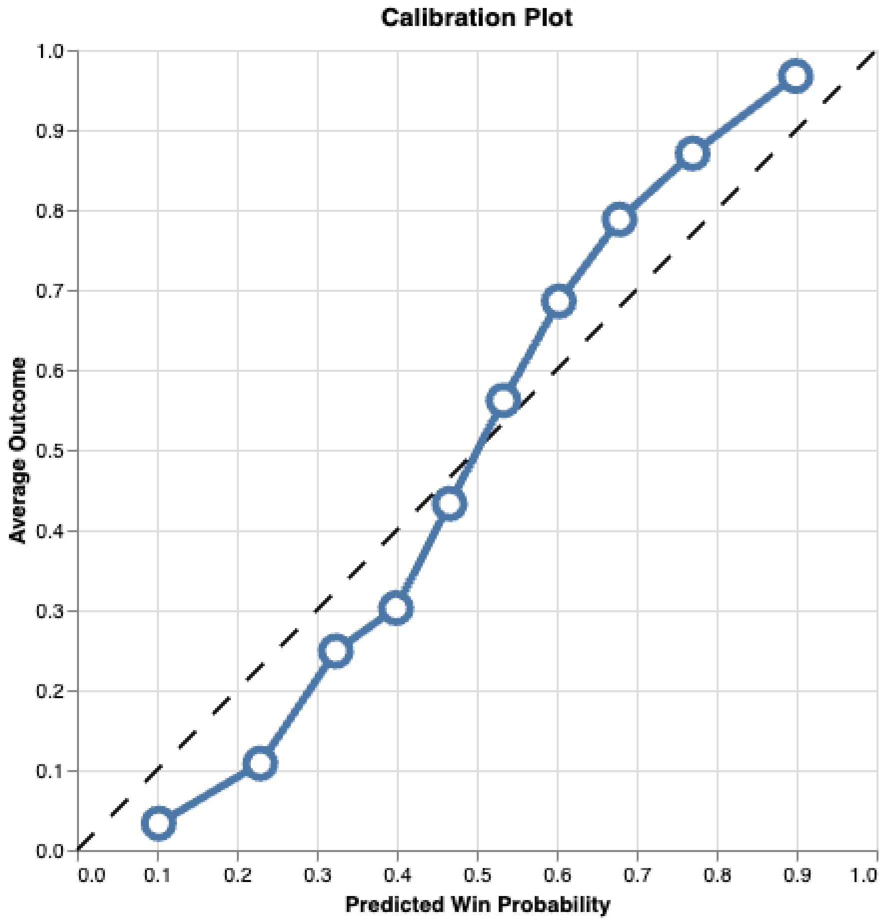


Calibration plot on 85777 games:



Further to-dos: A train test split on Warcraft 3 dataset

Calibration plot on 18370 games:



The plot is wider than the algo trained on 85777 game as above, however it's still closer to the horizontal line than the one from boxing games.

Then, we check the average games that a player has played.
For Warcraft 3 18370-game dataset, a player has played on average 38.83 games, compared to a boxer who has played about 3.2 games on average.

```
total_matches_df = games.winner.value_counts().add(games.loser.value_counts(), fill_value=0).sort_values()
✓ 0.0s

total_matches_df.describe()
✓ 0.0s

count     946.000000
mean       38.837209
std       110.627957
min         1.000000
25%         3.000000
50%         9.000000
75%        27.000000
max      1272.000000
Name: count, dtype: float64
```

Even after we focus on boxers who have played over 40 matches, we still have to include many boxers who have played few because they are the opponents of those frequently played boxers.

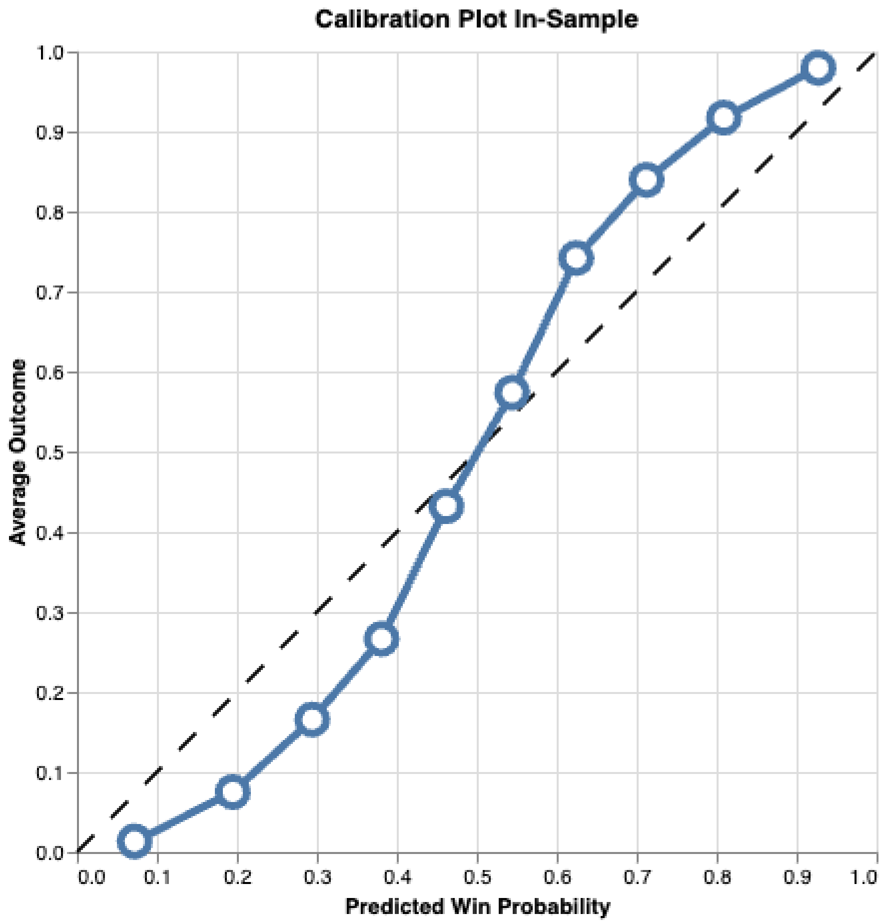
(see plots in page 5 on trueskill_boxers.pdf)

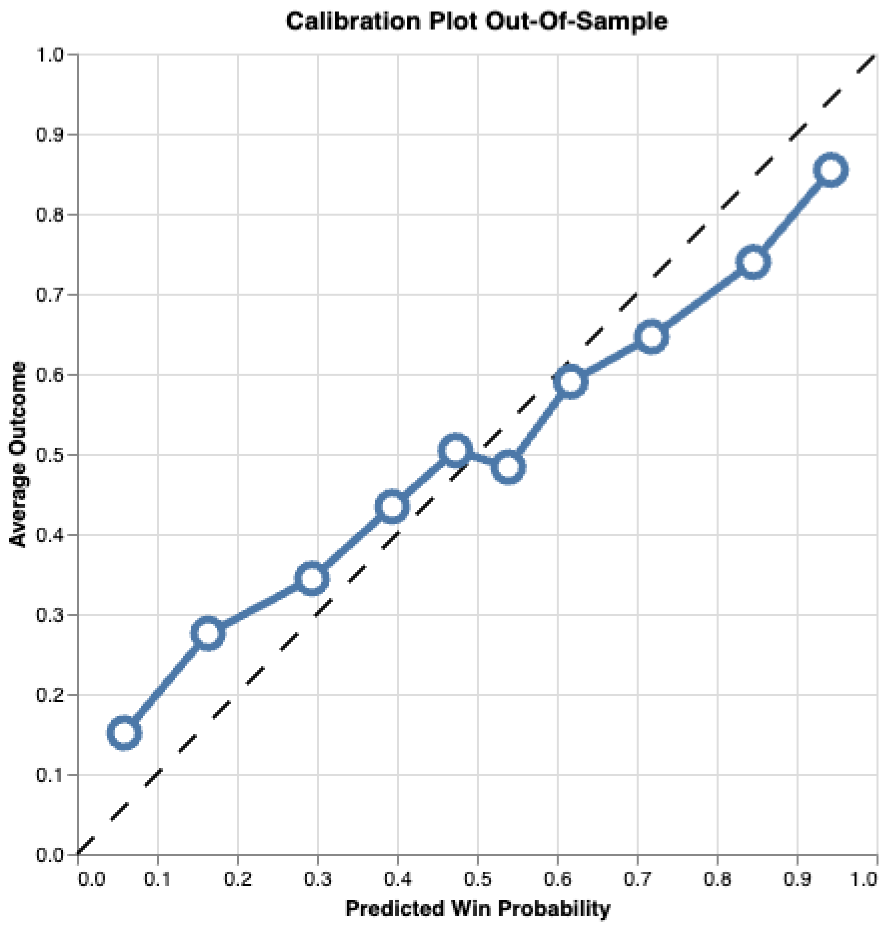
If we only look at the last 10000 games in warcraft3.csv whose competitor_1_score and competitor_2_score are both > -0.0001, there are 182 players who have played at least 40 matches in their career. The win rate for those players are quite high, which makes warcraft3 a very competitive game.

```
result_topplayer.win_rate.describe()
✓ 0.0s

count    181.000000
mean       0.485931
std        0.125252
min        0.162791
25%        0.407407
50%        0.499368
75%        0.576000
max        0.776119
Name: win_rate, dtype: float64
```

In-sample vs Out-of-sample calibration plot:
We first train a tstt algo on the training set, then for each player in the oos data, take the player's last available mu and sigma then compute win_prob. Finally, bucket the win_prob by 10 quantiles then plot against their corresponding actual win_outcome's mean to generate the calibration plot.





As in the Out-Of-Sample calibration plot, we still have a close to diagonal line for pred_win_prob vs avg_outcome. However, the model tends to underestimate the underdog player and overestimate the top dog since it only uses the historical data to calculate win_prob.