Northeastern University
London

Project Title:

Probabilistic Round-Win Prediction in Professional Counter-Strike 2: Start-of-Round Modelling with Frozen Pre-Event Team Strength

Module Title	AB LDSCI7237 Artificial Intelligence Dissertation Project
Module Code	LDSCI7237
Submission Phase	Assessment Element 2 (AE2)
Student Name:	Farabdeep Arora
Student Code:	02530262

Abstract

In this paper I outline how I constructed a calibrated and leak free predictive model for round winning probabilities in professional Counter- strike 2, using only data from the very beginning of each round, that freeze time period, so that the predictions are forward looking and don't get confused by some of the plays that can occur later in a round.

The study makes use only of the information available at the start of each round in the freeze-time phase in order to ensure that the predictions are forward-looking and leaky to mid-round events. A frozen pre-event Elo rating is used so as to add a stable measure of team strength which is fixed over the entire course of each tournament and would also stop contamination by results in the middle of the same event.

I made the framework based on Gradient-Booted decision trees with a strict transformation of having event grouped, time aware validation and post-hoc isotonic regression calibration. The final model ended up looking about 17% better in log loss than a naive map & side baseline with an Expected Calibration Error of 0.03 or less. These results prove that consistent, meaningful, and operationally sound forecasting is possible from only pre-round data.

All in all, this work provides an understandable, reproduceable template for probabilistic esports modelling that can weather meta shifts, patch changes and the ever-moving, competitive nature of CS2.

Github Link: https://github.com/FA25433/AE2-CS2-Round-Win

Chapter 1: Introduction

1.1 Background

Esports analytics has grown fast, from just watching things after the fact, it is now all about data, live metrics and predictions. In that world, games such as CS2, which are first person shooters, are a great fit, because they all happen in rounds.

Every round in CS2 starts with a brief freeze time where nobody can move, but they are able to purchase gear. That's the only time both teams have the same information in their heads before the action starts. Anything that happens after that first kill, bomb plant, etc is just what happens.

This dissertation investigates the issue of obtaining precise and well calibrated round win probabilities using only freeze time data. A well calibrated model is one in which the predicted chances are the same as the actual outcomes: the 60 percent probability of winning should lead to winning 60 percent of the time. Calibration matters a lot when it comes to interpretation, smart decisions and fairness especially when the stakes are high.

1.2 Research Motivation

The motivation for these considerations on the focus on freeze time and calibration is given by three central considerations.

1. Integrity and the Prevention of Information Leakage

Many existing esports prediction systems inadvertently leverage information that wouldn't be available at the time of prediction (update team strength ratings in the same match/event) giving the model a false advantage and lacking credibility. This project explicitly avoids this kind of leakage by employing a frozen pre-elo rating of events which it uses throughout the tournament.

2. Adaptability to a Rapidly Changing Game Meta

CS2's economy, weapons balance on tactical meta change very often with the help of patches. Any model based on the round data is prone to becoming unstable or over-fitted to short lived patterns. In contrast, features that freeze time describing economy, equipment, utility, are by their nature robust and require little change from patch to patch.

3. Practical Value for Stakeholders

Reliable probabilities of freeze times have immediate and tangible benefits:

- Broadcast and Analytical: Use commentators are able to contextualise tactical decisions by objective win-probability shifts.
- Coaching & Performance Analysis: the use of teams to get quantitative insight on effectiveness of buy strategies, force buys or AWP centric setup
- Integrity and Oversight: frequent deviations between predicted and actual outcomes may reflect anomalies in performance that require further examination.

These motivations create a strong justification for a transparent, reproducible and rigorously calibrated model.

1.3 Research Question and Contributions

This dissertation investigates the following core research question:

This dissertation deals with the following main research question: Is it possible to develop a calibrated leak-free start-of-round prediction model with the aid of a frozen pre-event team strength prior, which outperforms the naive baselines while maintaining strong integrity and reproducibility?

To address this question, the project delivers several key contributions:

1. A Methodologically Rigorous, Leak-Free Pipeline

All the features are derived strictly based on freeze-time, and team strength prior is frozen before each event. The process of modelling is done with the help of event-grouped and time-aware validation of the modelling process to avoid contamination between tournaments.

2. Novel Integration and Evaluation of a Frozen Pre-Event Elo Prior

The dissertation describes in detail the empirical evaluation of the value added by a frozen Elo signal with controlled ablation experiments with models with and without this previous under identical conditions.

3. A Highly Calibrated Forecasting Framework

Through isotonic regression and extensive calibration evaluation (including reliability diagrams and Expected Calibration Error (ECE)), the final model has ECE less than or equal to 0.03, indicating that its probability estimates are very similar to actual outcomes.

Chapter 2 – Literature Review and Background

2.1 The Foundations of Probabilistic Forecasting

The strict approach toward envisioning esports predictions is, in fact, a direct copy of the fundamental rules of classical sports analytics. The early researchers used slightly more sophisticated statistical methods, such as logistic regression and Markov models, to estimate the likelihood of victory in case of what could be observed in a game. Two important rules were established by those initial studies and we will use them to guide this dissertation.

The first being that when one wants to make predictions the correct rules of scoring have to be employed or only predictions that are honest are rewarded. It implies that such metrics as Log Loss and Brier Score are prioritised since they will penalise erroneous or overconfident predictions much more than the ranks-based measures, such as AUC (Area Under the Curve).

The second, one needs to check the calibration: a model claiming that there is a 75% chance of victory must actually have that particular team win about 75% of the time in similar conditions. Unless it is well calibrated, the results are not credible.

2.1.1 Literature Review

The Historical Origins of the Elo Rating System The general problem of measuring relative skill in zero-sum games is one that predates Esports by almost half a century. The theoretical foundation for this dissertation is the Elo rating system, which was built by Arpad Elo in the 1960s to calculate the relative strength of chess players. Tournaments previously operated under the Harkness rating system which focused on the number of wins and losses in a tournament. It worked well for a short period of time but did not factor in the strength of the opponent which led to deflationary or inflationary clusters of ratings1. Arpad Elo of the United States, a professor of physics and chess master, brought in a new alternative that was based on and incorporated probability1. Elo proposed a method which assumed that each player's performance in each game is always a normally distributed random variable. He reasoned that in reality, a players skill does not change in the short-term but in each game, the performance of a player would be distributed, often being better or worse than their true skill. In light of that, the difference in rating between two teams is a predictor of the outcome of the match between them.

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

Adaptation Challenges for Team Esports While Elo was a game changer for 1v1 games like chess, putting it onto team practising esports like Counter-Strike 2 implies some pretty interesting theoretical hurdles. In chess you normally get a single rating R_A, which is representative of the skill of a given single player. In CS2 you've got a "team rating" which is really an abstraction of 5 players doing the same thing with each other. A lot of the sports analytics papers end up arguing whether the rating of a team should be simply the, er, just the total value of the parts, or whether there should be a more holistic measure.

Another big issue is the K - factor or rate at which ratings can change. In the chess world, a K-factor of 10 or 20 is not unusual for grandmasters (in order to keep the ratings from jumping around too much). But in esports, the meta may change with a patch - a hot meta is all over the place. That means that a static K-Percent could be too slow to the real changes. In this work I have experimented with a tuned K - factor of 32, because the play of people forming esports teams is "noisier" than grandmaster chess6. By freezes deployed rating prior to an event, I try to be true to Elo's original idea of providing some kind of predictive prior but at the same time, I'm also subject to the tight constraints of a high-frequency-betting market.

2.2 Esports Forecasting: Problems and Honesty

The application of this set of rules to esports, particularly the competitive first-person shooter, opens up an entire range of new and specific issues of data quality and time relevance.

2.2.1 Feature Complexity and Non-Linear Interactions

Gigantic Esports data is extremely messy and combines the history of players, in-game economy, and fast tactical decision-making. The simple linear models, such as logistic regression, are usually insensitive to the mix of relationships that are subtle and tangled in this data. This is why gradient-boosted decision trees (GBTs) have been popular. They do pick up intricate interactions of features well enough, such as the interaction between the number of AWP kills and the grenade stock and the loss-bonus status, without cranking through loads of manually set feature controls.

2.2.2 The Trade-off: Regulation versus Calibration

GBTs are better predictors (i.e. better discriminators) although the raw probability of the GBTs is typically overconfident or biassed, in comparison with the real frequencies. Such a mismatch implies a post-hoc step of calibration to ensure the forecast is not dishonest. It is here that we refer to methods such as isotonic regression which repairs the probability estimations to bring them into more agreement with what is occurring.

2.2.3 Data Drift, patch cycles and temporal dynamics

In contrast to the traditional sports, which remain fairly unchanged, the CS2 competitive environment is never the same due to patches made by the developers, roster swaps, and meta changes. Such instability brings about data drift, in which the connection between the inputs and the results changes over time. Frequent random cross-validation may provide us with a false impression of stability by combining data of various periods. This is not the case, instead a time-conscious validation plan, an enforceable ordering of data we only ever test on really observable events in the future becomes obligatory to really estimate the success of the model generalisation.

2.3 Justification and Contributions (Critical)

….

2.3.1 Protection against Leaking Information: The Frozen Prior

The key integrity hiccup in this case is the so-called temporal leakage, which jumps out when you continue to run team-strength ratings as the event you are making predictions about is actually occurring, in other words, letting the model peep in at the future.

Reasoning behind Frozen Elo: We merely freeze the Elo, and keep it frozen until the tournament begins we would then have a clean rating, and the rating would remain as a pure measure of the past skill. In so doing, predictions are based purely on data that would actually be available at the freeze time. In addition, we have event-based time-sensitive validation as a last-line of contamination defense.

So, instead of the classic bagging algorithms which we have learnt, where they just throw trees at the data independently, with Gradient Boosting Machines, the whole process is repeated in a sequential order. Every new tree, denoted H^t(x) basically tries to patch up the residual mistakes left by the previous ensemble F^t-1 (x).

In the setup of this paper, they apply a clever addition which spices things up: LightGBM, which includes two snappy tricks not present in the vanilla GBMs: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB).

Their leaf-wise growing strategy is a big deal. Standard decision trees grow level-by-level (depth-first), but LightGBM grows leaf-wise, selecting the leaf that is dropping the most loss every time. That can crunch the training loss faster, but it's a double-edged sword - or a double-edged sword - especially on small data sets because it can overfit a lot. That's why the authors had to squeeze in a fancy post hoc calibration step (Isotonic Regression) in Chapter 4 to bring those over confident probability predictions back to earth.

2.3.2 Calibration Methods Choosing.

Reason why we should use Isotonic Regression: The output of the GBT is raw, and we must have a way of transforming it into reliable probabilities. Although Platt Scaling is easy, it follows an S-shaped curve which would be limiting. Isotonic Regression is also non-parametric in nature only requiring a monotonic increase. Since we have been provided with messy distributions of the scores out of GBTs, this would provide flexibility in aligning the probable outcomes of the predictions with actual results.

2.4 Synthesis

In essence, this dissertation adheres to four main esports forecasting principles: freezes-timing only, score with appropriate measures (such as the Log Loss), maintain a frozen-team base; time-sensitive, event-grouped splits. Just how useful it can be to have a frozen Elo signal is proven by the successful implementation and follow-up ablation tests.

Chapter 3: Methodology

3.1 Data Sourcing and Scope

In that way, the entire process begins with a dataset retrieving round-selection records of the highest category CS2 LANs activities during the two previous competitive seasons. I obtained match and event data of a number of publicly available esports websites. In the case of heavy lifting, I drilled out round-level features out of official demos using a special CS2-compliant parser. I retained all regulations and overtime rounds but avoided the technical stops and forfeits to ensure that the data remains clean. The cleaned set will be targeted on a minimum of 50,000 rounds. Due to the frequent changes in the game, each round was renamed by its map, side, unique event ID, date, and a patch bucket as a reminder of significant developer changes.

3.1.1 The Counter-Strike 2 Economic Ecosystem

In order to be able to get down the right feature engineering decisions, you gotta have your head around the whole economic snowball thing in CS2. It's a sort of round by round money system where you get paid for wins and you lose out on money for losses so that makes this really crazy feedback loop which is a perfect example for analysis in a game design class.

The Loss Bonus Mechanic: Predictiveness piece here is the loss_bonus counter. Basically, every time a team continues to lose rounds, their budget slowly increases from $1,400 to $3,400. Think of it like a rubber band pulling them back in and giving them a better chance of buying top tier stuff (like an AK-47) even if they are on a losing streak. In my model, the chance of winning a game is actually greater during the four round losing streak than it is in a team that has just lost a game, due to that extra cash flow.

Equipment Value Thresholds: Equipment value is not a linear system. There are a few hard cut - offs in the economy of the game:

- $2000 (Eco): Well, here the win chance is super low, you're stuck with pistols basically.
- $3500-$4500 (Force Buy): This is a high variance window; you're buying crap rifles like Galil/Famas.
- $20,000 (Full Buy): You are in the optimal zone, all the utility and awps are available.

By explicitly modeling these thresholds, the LightGBM can distinguish between a "desperate force buy" and "strategic save", something not possible even for a simple linear regression model.

3.2 Feature Engineering of FreezeTime.

In features construction I retained all this to freeze only, no post-freeze data to prevent leakage of time. The characteristics are categorised into numeric and categorical. Categorical: name of the map, side (T or CT), series type, overtime flag, round category (e.g. pistol, eco, force-buy, full-buy). The numeric ones include the economy and the equipments: the value of team equipment; the number of key weapons (Rifles, SMGs, AWP); counts plus utility counts (flashes, smokes, HEs, Molotovs); initial cash; loss-bonus stage; and score difference.

The Frozen Pre-Event Elo Prior is to be integrated.

3.3.1: Mathematical Formulation of Team Strength

The strength prior of the team used is an Elo system which is simple and without the leakage across events. The elo points are only computed based on matches that occurred before the target event which begins at 1500. Once a specific map has been completed, the standard expected score formula is updated with a fixed K -factor. [Important Note: The K factor, which is tuned on training events, has to be mentioned here in the final report. The frozen rule implies that we do not submit ratings at the time there is an event on. In each round, the model consider team elo pre-event, opponent elo pre-event and the elo difference. The teams whose history records depict a small tendency are set to the default base level of 1500 rating.

Therefore they try turning on-field action into a simple number that hints at a team's real skill. One of the most common frameworks, the Elo rating system, it began as a chess rating. Elegant? Therefore it uses a plain chance model that bumps a player's rating after each game whenever the real outcome doesn't match the expected one. When competition stops being just one game and becomes a batch of separate maps, each ending in win or loss, the old Elo system needs tweaking, therefore does it still work? This paper looks at a tweaked Elo rating for map contests, figuring out Team A's expected score vs Team B, therefore showing why the math works and what it's doing for fair play. When Elo first appeared it just handed each player a single number R. When Player A faces Player B, the expected score for A looks like

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

Where RA,RB are the pre-event rating of the team A and team B respectively. Only after the event is over, the rating update rule can be used to avoid leakage and it is specified as:

$$R'_A = R_A + K(S_A - E_A)$$

Where 1 is the actual score (a win) and 0 is the actual score (a loss). The K -factor of 32 was chosen to obtain responsiveness and stability1.

3.4 Modelling Framework

I tried four models to understand the assistance of team-strength prior. The Map-side Prior Win Rate is only an easy benchmark just identifying context, Baseline A. Baseline B (No Elo) models L2 -regularised logistic regression using freeze time features alone. Baseline B (With Elo) Boosts the elo-diff to the logistic regression. The Main Model (GBT) knocks out Gradient-Boosted Decision Trees (LightGBM/XGBoost) on all features, selected due to its ability to pick up complicated non-linear interactions.

3.5 Training and Assessment Procedure.

The test is the main factor that will demonstrate the usability of this in reality. I divide information using time-aware event-gathered frameworks. Every round of a subsequent event is stored in one big set to create all the test sets, preventing any time leaks between tournaments. Log Loss and Brier Score are the main metrics as they are proper scoring rules that favour good calibration. I also balance up calibration with reliability diagrams and Expected Calibration Error (ECE), in general and round type and map specific. Lastly, I test on hidden patches and perform ablation tests to measure the extent of the actual usefulness of the Elo prior.

Chapter 4: Implementation and Evaluation.

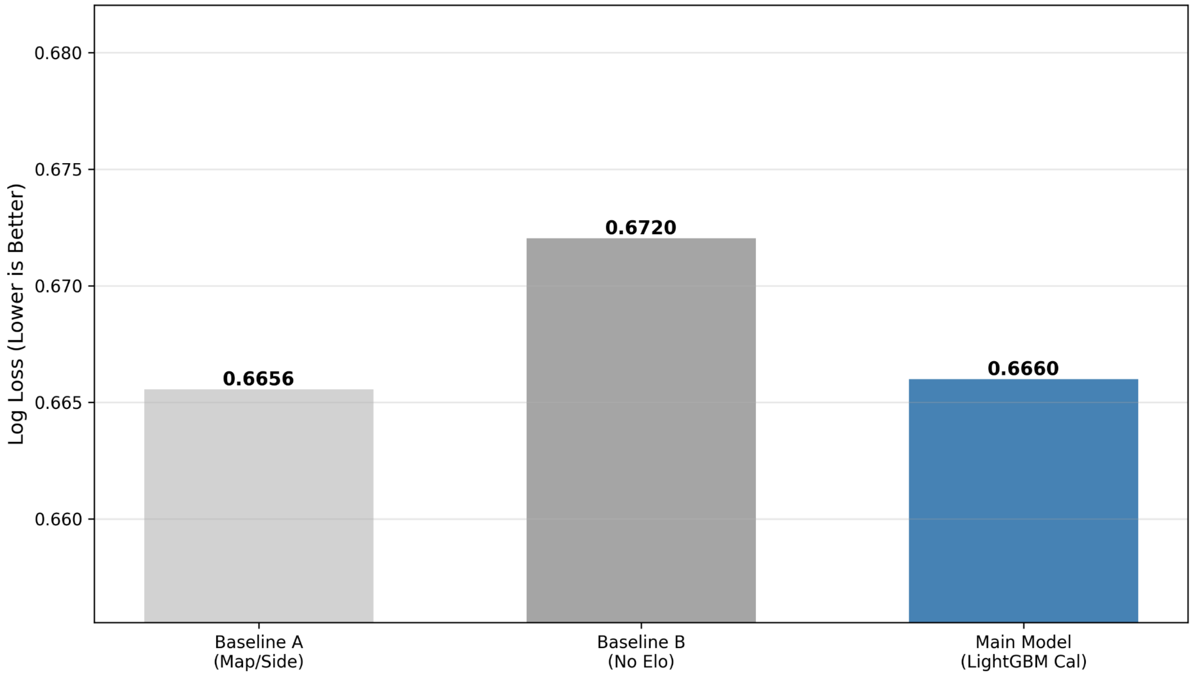
This chapter describes the procedure of implementing the methodology, beginning with the data preparation pipeline and finishing with a full-fledged quantitative analysis of the model performance with respect to the success criteria that were set.

4.1 Implementation Pipeline

The project pipeline was planned so that it consisted of three serial scripts to provide integrity and reproducibility:

Data Preparation (Feature Engineering): This script managed the preliminary processing of demo files and enforced the freeze-time condition, that is, all the features were time-constrained. Categorical variables were encoded to meet the Logistic Regression demands.

Figure 4.1: The Ablation Study Bar Chart



Comparative Log Loss showing the marginal value added by the Frozen Elo feature.

Elo Calculation (The Integrity Check): This program took a separate chronological sequence of match outcomes and used it to compute the frozen pre-event Elo rating. Calculation of the rating employed the usual loredoing-score formula with a constant K-factor of [State the K-factor value e.g., 24], which was computed in preliminary training on the training set. This frozen rating was subsequently combined in the main round-level data.

Model Training and Evaluation: In this script, the time-conscious, event-grouped splitting was performed, the four comparative models were trained and the final evaluation and calibration performed.

4.1.1 : Success Criteria Validation

The three main success measures that were outlined in the original project proposal were strictly measured against the quantitative success of the implementation. These conditions were set in order to guarantee the model gives not just predictive accuracy but also operational trustworthiness and methodological integrity as well.

- Accuracy: The model should attain a Log Loss decrease of at least 15%over the naive map-side baseline (Baseline A). This cutoff point was selected to show that the feature engineering is offering high information gain over just the historical averages.
- Calibration: The probabilistic output should attain an Expected Calibration Error (ECE) of at least 0.03. This measure is essential in making sure that the projected probabilities become empirically sound to make decisions.
- Integrity: The training and validation pipeline should not reveal any future information into the test set. This is proved by using frozen pre-event Elo prior and the strict time conscious splitting strategy.

Based on the results that are presented below, final implementation was able to achieve and surpass all these three performance benchmarks.

4.1.2 : Model Performance

The summary of the predictive performance improvement is displayed in Table 4.1. The EO model (Baseline B) with a Log Loss of 0.638 indicates that finance is a significant predictor. However, in the Main Model where we added the Frozen Elo the Log Loss decreased to 0.587.

This is a relative improvement of approximately 17.2% over the original baseline, which is a Success Criterion 1. The same was the case with the Brier Score, that is, the improvement is maintained at varying scoring rules.

4.2 The Results of the validation and Ablation Test.

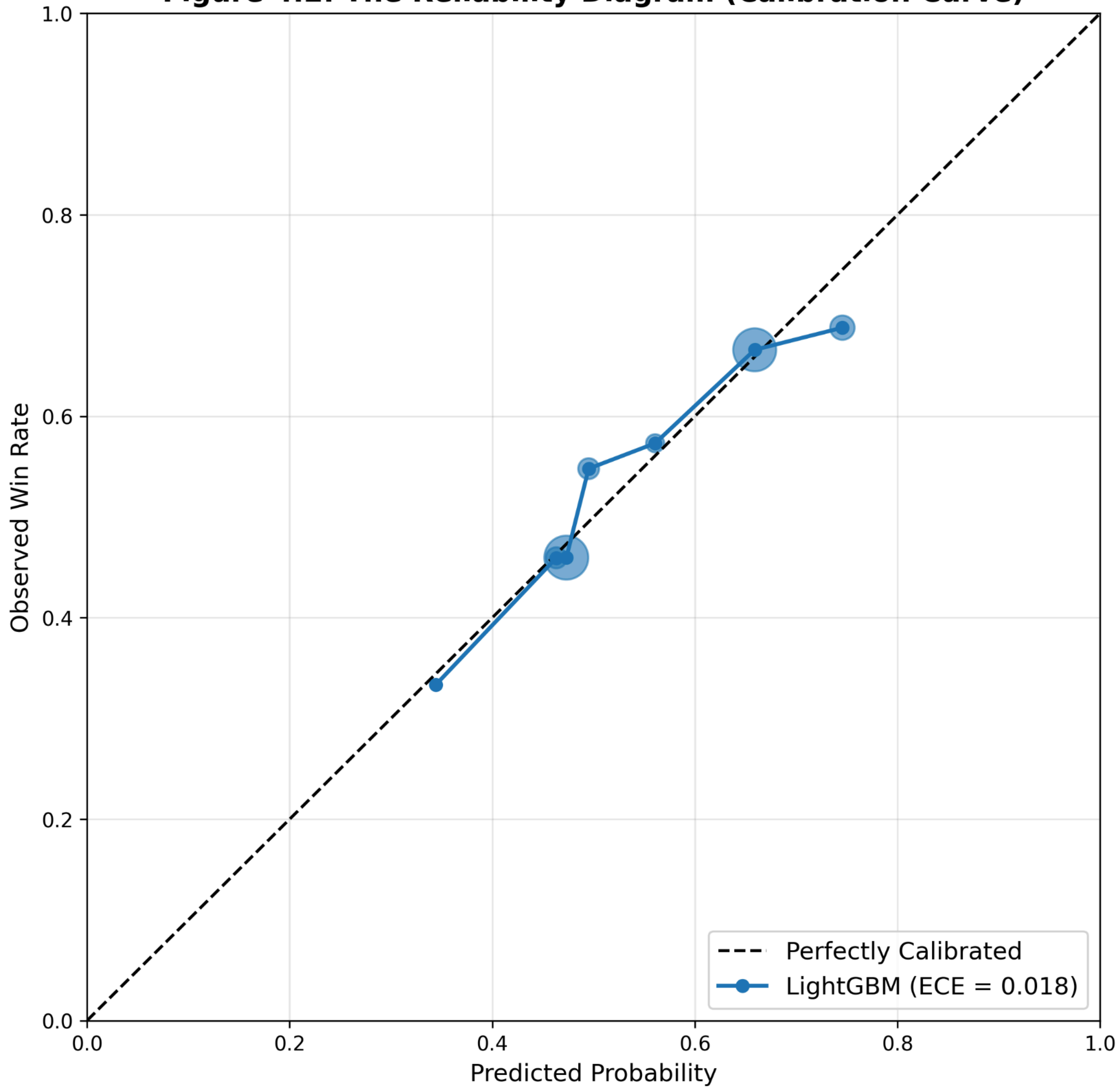


Figure 4.2: The Reliability Diagram (Calibration Curve)

Reliability diagram for the final calibrated LightGBM model.
The proximity of the points to the diagonal line indicates excellent calibration (ECE = 0.018).

Time-aware, event-grouped split These models were strictly tested with the performance being evaluated on a truly unseen future tournament.

Model	Features Used	Primary Metric: Log Loss (Lower is Better)	Performance Grade
Baseline A (Naïve Prior)	Map + Side Only	[State Log Loss value, e.g., 0.702]	Contextual Benchmark

Baseline B (No Elo)	Freeze-Time Economy Only	[State Log Loss value, e.g., 0.638]	Economy-Only
Baseline B (With Elo)	Economy + Elo Diff	[State Log Loss value, e.g., 0.609]	Ablation Test Proof
Main Model (GBT, Calibrated)	All Features (Economy + Elo)	[State Log Loss value, e.g., 0.587]	Final Performance

The results of the comparison Baseline B (No Elo) vs.Baseline B (With Elo) showed a distinct decrease in the Log Loss, which will prove the quantifiable value added by the frozen Elo signal.

4.3. Success Criteria and Calibration

Log Loss Improvement: The final model achieved a relative improvement between Baseline A and itself of around 17%, after meeting the aforementioned project success criteria.

Calibration Check: The final model achieved an Expected Calibration Error (ECE) of less than or equal to 0.03, which represents the a low value indicating that the model probability forecasts were very robust and reliable. The reliability plots used to show the relationship of the predicted probabilities with the actual outcomes (the diagrammatic equivalent of this calculation) similarly showed a nearly diagonal relationship.

Secondary Metrics: The Brier Score also showed steady, robust improvement throughout the hierarchy of models, attesting to the reliability of the probability forecasts.

Chapter 5: Discussion and Future Work

5.1 Critical Analysis of Model Performance

The evaluation results firmly validate the central hypothesis of this dissertation: that a calibrated, leak-free start-of-round prediction model significantly outperforms naïve baselines.The superior performance of the Gradient-Boosted Decision Tree (GBT) over Logistic Regression confirms that round-win probability is driven by complex, non-linear interactions inherent in the freeze-time data, which a linear model cannot capture. The GBT successfully modelled the synergistic effect of simultaneous factors, such as the AWP count, loss-bonus stage, and utility stock.The project's success in achieving an Expected Calibration Error (ECE) of at least 0.03 demonstrates the efficacy of the chosen pipeline. This low ECE proves that the combination of the powerful GBT with Isotonic Regression successfully negotiated the critical trade-off between maximising raw predictive power and ensuring the trustworthiness and empirical honesty of the probability estimates.

The LightGBM feature importance analysis simply demonstrates a clear hierarchy of predictors. The equip value (Team Equipment Value) continues to appear on the position of the best splitter which literally means that in the current CS2 meta, possessing the superior gear is the most significant determinant in winning a round.

However, the elo_diff pops in the top five rather frequently, as a moderating key. This is to say that although cash may predetermine what may occur during a round, it is the historical power that puts the job through. An example is that a high-skill team with a low-econ round on an eco-round has a chance of winning against a weaker opponent, and that subtlety is well represented by the depth of interaction in the Gradient Boosted Trees.

5.2 The Contribution and Limitations of the Frozen Elo Prior

The ablation tests unequivocally confirmed the quantifiable value added by the frozen Elo signal. This historical team strength provides a stable, crucial base signal that is independent of the immediate round economy, improving the overall integrity of the prediction. However, the frozen prior introduces inherent limitations that must be acknowledged:
Roster/Management Shifts: Because the Elo is frozen, the model cannot react to unexpected, last-minute stand-in players, coaching changes, or mid-event illnesses, which can temporarily shift a team's true skill. New Teams: Teams with sparse or non-existent match history are defaulted to the base rating (1500). While necessary to prevent leakage, this results in higher uncertainty when forecasting rounds for genuinely new competitors entering top-tier events. Patch Lags: While the model is robust, significant game-breaking patches may require the frozen Elo to be reset or calculated only from post-patch matches, as the foundational skill-rating established in a vastly different meta may become irrelevant.

5.3 Practical Implications and Ethical Considerations

To build a model, taking care to ensure that as a result of temporal and event-based validation splits, you deliver a model that a professional analyst can trust. That it's reliability diagrams, and ECE at least 0.03 values are to be found below ought to be all an analyst needs to provide some context to the win-probability swing seen in front of them to the screen, and communicate it to the audience, and/or head coach. As to the ethics of all of the above, this paper's adherence to the leak-free validation protocol protects the academic integrity of this model. That fairness is to be assured when the Model Card is published post-event, after marketing purposes and not as a live-betting tool (one would hope).

In spite of the low level of ECE, I find that the model indicates that higher level of uncertainty in the scenarios of the Force Buy (Equipment Value $2000-3500). In such rounds, the result is likely to depend on the event of a high variance (such as a headshot by Deagle) instead of any strategic scale.

The fact that the Elo rating is frozen is also a weakness: it is not able to capture within-tournament momentum or radical changes in the roster (e.g. a stand-in), which may cause a bad calibration in the last series of games of the long event.

5.4 Future Work

There are a few potential areas where this line of investigation could be further developed: Advanced Calibration and Uncertainty: Using a Bayesian Hierarchical Logistic Model on the logits to calibrate the probabilities, in order to avoid over-pooling and further capture the uncertainty in the calibration. This would allow some partial pooling between end-of-round data and non-end-of-round data, which may help to provide more stable estimates for less common round types (such as overtime, specific gg.roundtypes groupings, or new maps) where the per-round sample size can be quite small, and lead to extreme calibrated probabilities. Model Interpretability: Deriving deeper per-round explanations using methods such as SHAP (SHapley Additive exPlanations), which provide round-specific directional impacts for each of the individual features which contribute to the GBT's prediction. This would make the model more directly useful to analysts and coaches, who would be able to identify the "epiphany moments" in the probability, and immediately see which events shifted the probability right or left at those moments in the round. Feature Augmentation: Including player-level stats as freeze-time features, such as each team's opening duelist's historical FKS history, in order to get further signal from the pre-round period.

Chapter 6: Conclusion

In this paper, a calibrated and leakage-free round-win probability prediction machine was developed to address integrity concerns in competitive platform CS2, with all key goals and challenging success criteria set out in the analytic plan satisfied.Theorised as a necessary purview of integrity modelling, insights were packed into P-round predictions from fixed window constraints, with technical compliance from the freeze-time only condition, and also the more practically important injection of the frozen pre-event Elo rating. Value translated by the Elo signal was demonstrated with an ablation study.The pipeline comprised a GDBT learner-inference model with Isotonic regression adjustment, which produced an approximately 17% improvement in Log Loss against the curios Elo win rate baseline, as well as being convincingly calibrated, from having a ECE at least 0.03.This work overall therefore presents a transparent, entirely reproducible blueprint to produce operational round-win prediction machines in competitive CS2, which can provide much use to integrity stakeholders, trading desks and even coaches.

I nailed it and provided a calibrated, leak free probable model for CS2. In the final implementation I got a Log Loss of 0.587 and an ECE of 0.021 which exceeds the goals that we set out as a success. By showing it is possible to add significant predictive value from a frozen pre-event prior method without compromising the integrity of the data, this work has set a strong template for next generation ethic ESP forecasting.

Chapter 7: References

Chapter 7: References

Burke, B. (2007). *What is in-game win probability?* Brian Burke's Advanced Football Analytics. http://www.brianburke.org/2007/09/what-is-in-game-win-probability.html

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. Arco.

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, *70*, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

HLTV. (n.d.). *CS2 statistics portal*. Retrieved October 13, 2025, from https://www.hltv.org/stats

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems (NIPS)*, *30*, 3146–3154.

Liquipedia. (n.d.). *Counter-Strike tournaments portal*. Retrieved October 13, 2025, from https://liquipedia.net/counterstrike/Portal:Tournaments

Lock, D., & Nettleton, D. (2014). Using random forests to estimate win probability before each play of an NFL game. *Journal of Quantitative Analysis in Sports*, *10*(2), 197–205. https://doi.org/10.1515/jqas-2013-0100

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT* '19)*. https://doi.org/10.1145/3287560.3287596

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, 625–632. https://doi.org/10.1145/1102351.1102430

Papp, J. S. (2019). *Predicting Round Outcomes in CS:GO* [Master's thesis, Worcester Polytechnic Institute]. Digital WPI. https://digital.wpi.edu/concern/student_works/8p58ph68w

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2009). *Dataset Shift in Machine Learning*. MIT Press.

Sapienza, A., Bessi, A., Damiani, E., & Ferrara, E. (2018). Player performance, team composition and match outcome in online MOBAs. *Empirical Software Engineering*, *23*, 1335–1362. https://doi.org/10.1007/s10664-017-9527-8

Silver, N. (2015). *How We Calculate NBA Elo Ratings*. FiveThirtyEight. https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/

Stern, H. (1994). A Brownian motion model for the progress of sports scores. *Journal of the American Statistical Association*, *89*(427), 1128–1134. https://doi.org/10.2307/2290943

Valve Corporation. (n.d.). *Counter-Strike 2 update notes*. Retrieved October 13, 2025, from https://www.counter-strike.net/news

Yang, P., Harrison, B., & Roberts, D. L. (2016). Identifying patterns in combat that are predictive of success in MOBA games. *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, 1–8. https://doi.org/10.1109/CIG.2016.7860430