A Meta-Analysis of Overfitting in Machine Learning

Rebecca Roelofs[*]
UC Berkeley
`roelofs@berkeley.edu`

Sara Fridovich-Keil[*]
UC Berkeley
`sfk@berkeley.edu`

John Miller
UC Berkeley
`miller_john@berkeley.edu`

Vaishaal Shankar
UC Berkeley
`vaishaal@berkeley.edu`

Moritz Hardt
UC Berkeley
`hardt@berkeley.edu`

Benjamin Recht
UC Berkeley
`brecht@berkeley.edu`

Ludwig Schmidt
UC Berkeley
`ludwig@berkeley.edu`

Abstract

We conduct the first large meta-analysis of overfitting due to test set reuse in the machine learning community. Our analysis is based on over one hundred machine learning competitions hosted on the Kaggle platform over the course of several years. In each competition, numerous practitioners repeatedly evaluated their progress against a holdout set that forms the basis of a public ranking available throughout the competition. Performance on a separate test set used only once determined the final ranking. By systematically comparing the public ranking with the final ranking, we assess how much participants adapted to the holdout set over the course of a competition. Our study shows, somewhat surprisingly, little evidence of substantial overfitting. These findings speak to the robustness of the holdout method across different data domains, loss functions, model classes, and human analysts.

1 Introduction

The holdout method is central to empirical progress in the machine learning community. Competitions, benchmarks, and large-scale hyperparameter search all rely on splitting a data set into multiple pieces to separate model training from evaluation. However, when practitioners repeatedly reuse holdout data, the danger of overfitting to the holdout data arises [6, 13].

Despite its importance, there is little empirical research into the manifested robustness and validity of the holdout method in practical scenarios. Real-world use cases of the holdout method often fall outside the guarantees of existing theoretical bounds, making questions of validity a matter of guesswork.

Recent replication studies [16] demonstrated that the popular CIFAR-10 [10] and ImageNet [5, 18] benchmarks continue to support progress despite years of intensive use. The longevity of these benchmarks perhaps suggests that overfitting to holdout data is less of a concern than reasoning from first principles might have suggested. However, this is evidence from only two, albeit important, computer vision benchmarks. It remains unclear whether the observed phenomenon is specific to the data domain, model class, or practices of vision researchers. Unfortunately, these replication studies

---

[*] Equal contribution

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

required assembling new test sets from scratch, resulting in a highly labor-intensive analysis that is difficult to scale.

In this paper, we empirically study holdout reuse at a significantly larger scale by analyzing data from 120 machine learning competitions on the popular Kaggle platform [2]. Kaggle competitions are a particularly well-suited environment for studying overfitting since data sources are diverse, contestants use a wide range of model families, and training techniques vary greatly. Moreover, Kaggle competitions use public and private test data splits which provide a natural experimental setup for measuring overfitting on various datasets.

To provide a detailed analysis of each competition, we introduce a coherent methodology to characterize the extent of overfitting at three increasingly fine scales. Our approach allows us both to discuss the overall "health" of a competition across all submissions and to inspect signs of overfitting separately among the top submissions. In addition, we develop a statistical test specific to the classification competitions on Kaggle to compare the submission scores to those arising in an ideal null model that assumes no overfitting. Observed data that are close to data predicted by the null model is strong evidence against overfitting.

Overall, we conclude that the classification competitions on Kaggle show little to no signs of overfitting. While there are some outlier competitions in the data, these competitions usually have pathologies such as non-i.i.d. data splits or (effectively) small test sets. Among the remaining competitions, the public and private test scores show a remarkably good correspondence. The picture becomes more nuanced among the highest scoring submissions, but the overall effect sizes of (potential) overfitting are typically small (e.g., less than 1% classification accuracy). Thus, our findings show that substantial overfitting is unlikely to occur naturally in regular machine learning workflows.

2 Background and setup

Before we delve into the analysis of the Kaggle data, we briefly define the type of overfitting we study and then describe how the Kaggle competition format naturally lends itself to investigating overfitting in machine learning competitions.

2.1 Adaptive overfitting

"Overfitting" is often used as an umbrella term to describe any unwanted performance drop of a machine learning model. Here, we focus on *adaptive* overfitting, which is overfitting caused by test set reuse. While other phenomena under the overfitting umbrella are also important aspects of reliable machine learning (e.g. performance drops due to distribution shifts), they are beyond the scope of our paper since they require an experimental setup different from ours.

Formally, let $f : \mathcal{X} \to \mathcal{Y}$ be a trained model that maps examples $x \in \mathcal{X}$ to output values $y \in \mathcal{Y}$ (e.g., class labels or regression targets). The standard approach to measuring the performance of such a trained model is to define a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ and to draw samples $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ from a data distribution $\mathcal{D}$ which we then use to evaluate the test loss $L_S(f) = \sum_{i=1}^{n} L(f(x_i), y_i)$. As long as the model f does not depend on the test set S, standard concentration results [19] show that $L_S(f)$ is a good approximation of the true performance given by the population loss $L_{\mathcal{D}}(f) = \mathbb{E}_{\mathcal{D}}[L(f(x), y)]$.

However, machine learning practitioners often undermine the assumption that f does not depend on the test set by selecting models and tuning hyperparameters based on the test loss. Especially when algorithm designers evaluate a large number of different models on the same test set, the final classifier may only perform well on the specific examples in the test set. The failure to generalize to the entire data distribution $\mathcal{D}$ manifests itself in a large *adaptivity gap* $L_{\mathcal{D}}(f) - L_S(f)$ and leads to overly optimistic performance estimates.

2.2 Kaggle

Kaggle is the most widely used platform for machine learning competitions, currently hosting 1,461 active and completed competitions. Various organizations (companies, educators, etc.) provide the datasets and evaluation rules for the competitions, which are generally open to any participant. Each competition is centered around a dataset consisting of a training set and a test set.

Considering the danger of overfitting to the test set in a competitive environment, Kaggle subdivides each test set into public and private components. The subsets are randomly shuffled together and the entire test set is released without labels, so that participants should not know which test samples belong to which split. Hence participants submit predictions for the *entire* test set. The Kaggle server then internally evaluates each submission on both public and private splits and updates the public competition leaderboard only with the score on the public split. At the end of the competition, Kaggle releases the private scores, which determine the winner.

Kaggle has released the MetaKaggle dataset[2], which contains detailed information about competitions, submissions, etc. on the Kaggle platform. The structure of Kaggle competitions makes MetaKaggle a useful dataset for investigating overfitting empirically at a large scale. In particular, we can view the public test split S_{public} as the regular test set and use the held-out private test split S_{private} to approximate the population loss. Since Kaggle competitors do not receive feedback from S_{private} until the competition has ended, we assume that the submitted models may be overfit to S_{public} but not to S_{private}.[3] Under this assumption, the difference between private and public loss $L_{S_{\text{private}}}(f) - L_{S_{\text{public}}}(f)$ is an approximation of the adaptivity gap $L_{\mathcal{D}}(f) - L_S(f)$. Hence our setup allows us to estimate the amount of overfitting occurring in a typical machine learning competition. In the rest of this paper, we will analyze the public versus private score differences as a proxy for adaptive overfitting.

Due to the large number of competitions on the Kaggle platform, we restrict our attention to the most popular classification competitions. In particular, we survey the competitions with at least 1,000 submissions before the competition deadline. Moreover, we include only competitions with evaluation metrics that have at least 10 competitions. These metrics are classification accuracy, AUC (area under curve), MAP@K (mean average precision), the logistic loss, and a multiclass variant of the logistic loss. Appendix A provides more details about our selection criteria.

3 Detailed analysis of competitions scored with classification accuracy

We begin with a detailed look at classification competitions scored with the standard accuracy metric. Classification is the prototypical machine learning task and accuracy is a widely understood performance measure. This makes the corresponding competitions a natural starting point to understand nuances in the Kaggle data. Moreover, there is a large number of accuracy competitions, which enables us to meaningfully compare effect sizes across competitions. As we will see in Section 3.3, accuracy competitions also offer the advantage that we can obtain measures of statistical uncertainty. Later sections will then present an overview of the competitions scored with other metrics.

We conduct our analysis of overfitting at three levels of granularity that become increasingly stringent. The first level considers all submissions in a competition and checks for systematic overfitting that would affect a substantial number of submissions (e.g., if public and private score diverge early in the competition). The second level then zooms into the top 10% of submissions (measured by public accuracy) and conducts a similar comparison of public to private scores. The goal here is to understand whether there is more overfitting among the best submissions since they are likely most adapted to the test set. The third analysis level then takes a mainly quantitative approach and computes the probabilities of the observed public vs. private accuracy differences under an ideal null model. This allows us to check if the observed gaps are larger than purely random fluctuations.

In the following subsections, we will apply these three analysis methods to investigate four accuracy competitions. These four competitions are the accuracy competitions with the largest number of submissions and serve as representative examples for a typical accuracy competition in the MetaKaggle dataset (see Table 1 for information about these competitions). Section 3.4 then complements these analyses with a quantitative look at all competitions before we summarize our findings for accuracy competitions in Section 3.5.

---

[2] `https://www.kaggle.com/kaggle/meta-kaggle`

[3] Since test examples without labels are available, contestants may still overfit to S_{private} using an unsupervised approach. While such overfitting may have occurred in a limited number of competitions, we base our analysis on the assumption that unsupervised overfitting did not occur widely with effect sizes that are large compared to overfitting to the public test split.

Table 1: The four accuracy competitions with the largest number of submissions. n_{public} is the size of the public test set and n_{private} is the size of the private test set.

ID	Name	# Submissions	n_{public}	n_{private}
5275	Can we predict voting outcomes?	35,247	249,344	249,343
3788	Allstate Purchase Prediction Challenge	24,532	59,657	139,199
7634	TensorFlow Speech Recognition Challenge	24,263	3,171	155,365
7115	Cdiscount's Image Classification Challenge	5,859	53,0455	1,237,727

3.1 First analysis level: visualizing the overall trend

As mentioned in Section 2.2, the main quantities of interest are the accuracies on the public and private parts of the test set. In order to visualize this information at the level of all submissions to a competition, we create a scatter plot of the public and private accuracies. In an ideal competition with a large test set and without any overfitting, the public and private accuracies of a submission would all be almost identical and lie near the $y = x$ diagonal. On the other hand, substantial deviations from the diagonal would indicate gaps between the public and private accuracy and present possible evidence of overfitting in the competition.

Figure 1 shows such scatter plots for the four accuracy competitions mentioned above. In addition to a point for each submission, the scatter plots also contain a linear regression fit to the data. All four plots show a linear fit that is close to the $y = x$ diagonal. In addition, three of the four plots show very little variation around the diagonal. The variation around the diagonal in the remaining competition is largely symmetric, which indicates that it is likely the effect of random chance.

These scatter plots can be seen as indicators of overall competition "health": in case of pervasive overfitting, we would expect a plateauing trend where later points mainly move on the x-axis (public accuracy) but stagnate on the y-axis (private accuracy). In contrast, the four plots in Figure 1 show that as submissions progress on the public test set, they see corresponding improvements also on the private test set. Moreover, the public scores remain representative of the private scores.

As can be seen in Appendix B.3, this overall trend is representative for the 34 accuracy competitions. All except one competition show a linear fit close to the main diagonal. The only competition with a substantial deviation is the "TAU Robot Surface Detection" competition (ID 12598). We contacted the authors of this competition and confirmed that there are subtleties in the public / private split which undermine the assumption that the two splits are i.i.d. Hence we consider this competition to be an outlier since it does not conform to the experimental setup described in Section 2.2. So at least on the coarse scale of the first analysis level, there are little to no signs of adaptive overfitting: it is easier to make genuine progress on the data distribution in these competitions than to substantially overfit to the test set.

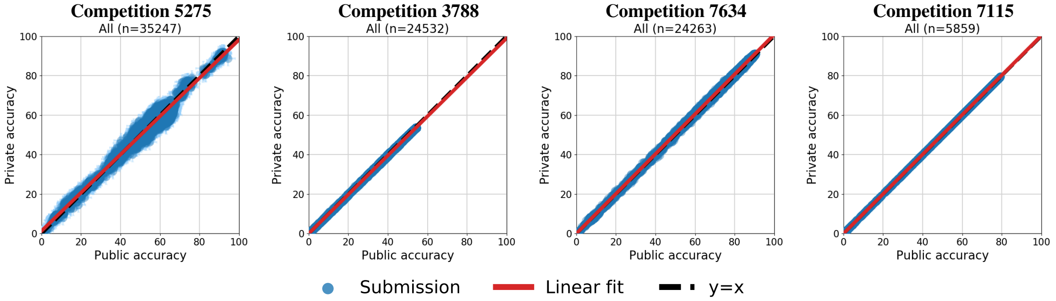


Figure 1: Private versus public accuracy for all submissions in the most popular Kaggle accuracy competitions. Each point corresponds to an individual submission (shown with 95% Clopper-Pearson confidence intervals, although the confidence intervals are smaller than the plotted data points).

3.2 Second analysis level: zooming in to the top submissions

While the scatter plots discussed above give a comprehensive picture of an entire competition, one concern is that overfitting may be more prevalent among the submissions with the highest public accuracy since they may be more adapted to the public test set. Moreover, the best submissions are those where overfitting would be most serious since invalid accuracies there would give a misleading impression of performance on future data. So to analyze the best submissions in more detail, we also created scatter plots for the top 10% of submissions (as scored by public accuracy).

Figure 2 shows scatter plots for the same four competitions as before. Since the axes now encompass a much smaller range (often only a few percent), they give a more nuanced picture of the performance among the best submissions.

In the leftmost and rightmost plot, the linear fit for the submissions still closely tracks the $y = x$ diagonal. Hence there is little sign of overfitting also among the top 10% of submissions. On the other hand, the middle two plots in Figure 2 show noticeable deviations from the main diagonal. Interestingly, the linear fit is *above* the diagonal in Competition 7634 and *below* the main diagonal in Competition 3788. The trend in Competition 3788 is more concerning since it indicates that the public accuracy overestimates the private accuracy. However, in both competitions the absolute effect size (deviation from the diagonal) is small (about 1%). It is also worth noting that the accuracies in Competition 3788 are not in a high-accuracy regime but around 55%. So the relative error from public to private test set is small as well.

Appendix B.4 contains scatter plots for the remaining accuracy competitions that show a similar overall trend. Besides the Competition 12598 discussed in the previous subsection, there are two additional competitions with a substantial public vs. private deviation. One competition is #3641, which has a total test set size of about 4,000 but is only derived from 7 human test subjects (the dataset consists of magnetoencephalography (MEG) recordings from these subjects). The other competition is 12681, which contains a public test set of size 90. Very small (effective) test sets make it easier to reconstruct the public / private split (and then to overfit), and also make the public and private scores more noisy. Hence we consider these two competitions to be outliers. Overall, the second analysis level shows that if overfitting occurs among the top submissions, it only does so to a small extent.

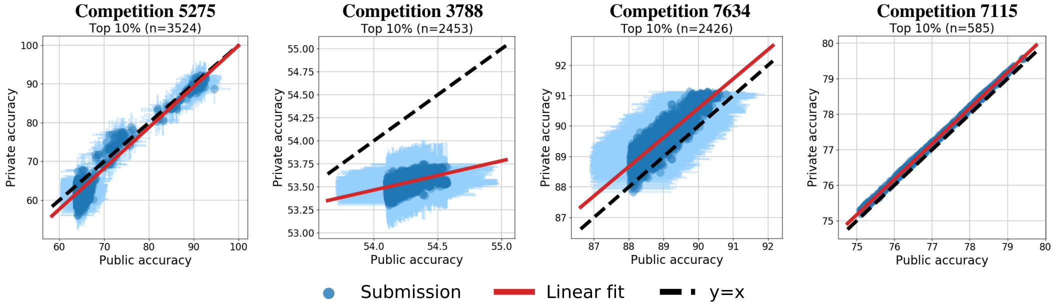


Figure 2: Private versus public accuracy for the top 10% of submissions in the most popular Kaggle accuracy competitions. Each point corresponds to an individual submission (shown with 95% Clopper-Pearson confidence intervals).

3.3 Third analysis level: quantifying the amount of random variation

When discussing deviations from the ideal $y = x$ diagonal in the previous two subsections, an important question is how much variation we should expect from random chance. Due to the finite sizes of the test sets, the public and private accuracies will never match exactly and it is a priori unclear how much of the deviation we can attribute to random chance and how much to overfitting.

To quantitatively understand the expected random variation, we compute the probability of a given public vs. private deviation (*p-value*) for a simple null model. By inspecting the distribution of the resulting p-values, we can then investigate to what extent the observed deviations can be attributed to random chance.

We consider the following null model under which we compute p-values for observing certain gaps between the public and private accuracies. We fix a submission that makes a given number of mistakes on the *entire* test set (public and private split combined). We then randomly split the test set into two parts with sizes corresponding to the public and private splits of the competition. This leads to a certain number of mistakes (and hence accuracy) on each of the two parts. The p-value for this submission is then given by the probability of the event

$$|\text{public\_accuracy} - \text{private\_accuracy}| \geq \varepsilon \tag{1}$$

where ε is the observed deviation between public and private accuracy. We describe the details of computing these p-values in Appendix B.5.

Figure 3 plots the distribution of these p-values for the same four competitions as before. To see potential effects of overfitting, we show p-values for all submissions, the top 10% of submissions (as in the previous subsection), and for the first submission for each team in the competition. We plot the first submissions separately since they should be least adapted to the test set and hence show the smallest amount of overfitting.

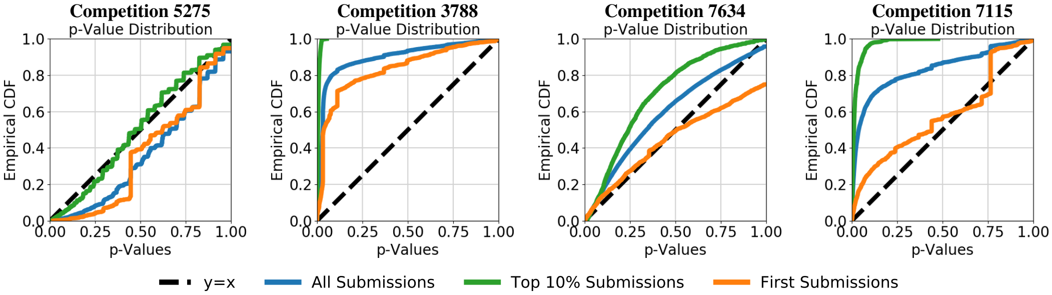


Figure 3: CDFs of the p-values for three (sub)sets of submissions in the four accuracy competitions with the largest number of submissions.

Under our ideal null hypothesis, the p-values would have a uniform distribution with a CDF following the $y = x$ diagonal. However, this is only the case for Competition 5275 and the first submissions in Competitions 7634 & 7115, and even there only approximately. Most other curves show a substantially larger number of small p-values (large gaps between public and private accuracy) than expected under the null model. Moreover, the top 10% of submissions always exhibit the smallest p-values, followed by all submissions and then the first submissions.

When interpreting these p-value plots, we emphasize that the null model is highly idealized. In particular, we assume that every submission is evaluated on its own *independent* random public / private split, which is clearly not the case for Kaggle competitions. So for correlated submissions, we would expect clusters of approximately equal p-values that are unlikely to arise under the null model. Since it is plausible that many models are trained with standard libraries such as XGBoost [4] or scikit-learn [14], we conjecture that correlated models are behind the jumps in the p-value CDFs.

In addition, it is important to note that for large test sets (e.g., the 198,856 examples in Competition 3788), even very small systematic deviations between the public and private accuracies are statistically significant and hence lead to small p-values. So while the analysis based on p-value plots does point towards irregularities such as overfitting, the overall effect size (deviation between public and private accuracies) can still be small.

Given the highly discriminative nature of the p-values, a natural question is whether any competition exhibits approximately uniform p-values. As can be seen in Appendix B.5, some competitions indeed have p-value plots that are close to the uniform distribution under null model (Figure 11 in the same appendix highlights four examples). Due to the idealized null hypothesis, this is strong evidence that these competitions are free from overfitting.

3.4 Aggregate view of the accuracy competitions

The previous subsections provided tools for analyzing individual competitions and then relied on a qualitative survey of all accuracy competitions. Due to the many nuances and failure modes in machine learning competitions, we believe that this case-by-case approach is the most suitable for the Kaggle data. But since this approach also carries the risk of missing overall trends in the data, we complement it here with a quantitative analysis of all accuracy competitions.

In order to compare the amount of overfitting across competitions, we compute the *mean accuracy difference* across all submissions in a competition. Specifically, let $\mathcal{C}$ be a set of submissions for a given competition. Then the mean accuracy difference of the competition is defined as

$$\text{mean\_accuracy\_difference} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \text{public\_accuracy}(i) - \text{private\_accuracy}(i)\ . \qquad (2)$$

A larger mean accuracy difference indicates more potential overfitting. Note that since accuracy is in a sense the opposite of loss, our computation of mean accuracy difference is the opposite of our earlier expression for the generalization gap.

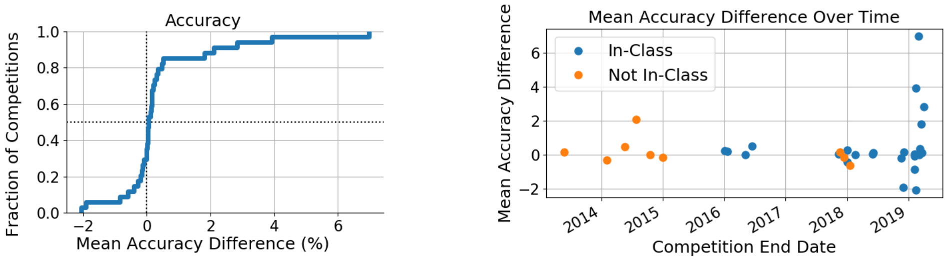


Figure 4: Left: Empirical CDF of the mean accuracy differences (%) for 34 accuracy competitions with at least 1,000 submissions. Right: Mean accuracy differences versus competition end date for the same competitions.

Figure 4 shows the empirical CDF of the mean accuracy differences for all accuracy competitions. While the score differences between -2% to $+2\%$ are approximately symmetric and centered at zero (as a central limit theorem argument would suggest), the plot also shows a tail with larger score differences consisting of five competitions. Figure 8 in Appendix B.2 aggregates our three types of analysis plots for these competitions.

We have already mentioned three of these outliers in the discussion so far. The worst outlier is Competition 12598 which contains a non-i.i.d. data split (see Section 3.1). Section 3.2 noted that Competitions 3641 and 12681 had very small test sets. Similarly, the other two outliers (#12349 and #5903) have small private test sets of size 209 and 100, respectively. Moreover, the public - private accuracy gap *decreases* among the very best submissions in these competitions (see Figure 8 in Appendix B.2). Hence we do not consider these competitions as examples of adaptive overfitting.

As a second aggregate comparison, the right plot in Figure 4 shows the mean accuracy differences vs. competition end date and separates two types of competitions: in-class and other, which are mainly the “featured” competitions on Kaggle with prize money. Interestingly, many of the competitions with large public - private deviations are from 2019. Moreover, the large deviations are almost exclusively in-class competitions. This may indicate that in-class competitions undergo less quality control than the featured competitions (e.g., have smaller test sets), and that the quality control standards on Kaggle may change over time.

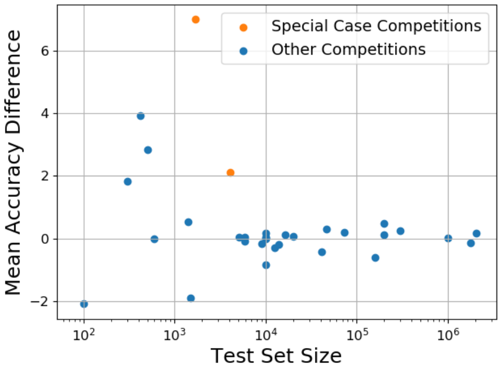


Figure 5: Mean accuracy differences versus test set size (public and private combined) for 32 accuracy competitions with at least 1,000 submissions and available test set size (the test set sizes for two competitions with at least 1,000 submissions were not available from the MetaKaggle dataset).

As a third aggregate comparison, Figure 5 shows the mean accuracy differences vs. test set sizes, with orange dots to indicate competitions we flagged as having either a known or a non-i.i.d. public vs. private test set split. Although a reliable recommendation for applied machine learning will require broader investigation, our results for accuracy competitions suggest that at least 10,000 examples is a reasonable minimum test set size to protect against adaptive overfitting.

3.5 Did we observe overfitting?

The preceding subsections provided an increasingly fine-grained analysis of the Kaggle competitions evaluated with classification accuracy. The scatter plots for all submissions in Section 3.1 show a good fit to the $y = x$ diagonal with small and approximately symmetric deviations from the diagonal. This is strong evidence that the overall competition is not affected by substantial overfitting. When restricting the plots to the top submissions in Section 3.2, the picture becomes more varied but the largest overfitting effect size (public - private accuracy deviation) is still small. Thus for some competitions we observe evidence of mild overfitting, but always with small effect size. In both sections we have identified outlier competitions that do not follow the overall trend but also have issues such as small test sets or non-i.i.d. splits. At the finest level of our analysis (Section 3.3), the p-value plots show that the data is only sometimes in agreement with an idealized null model for no overfitting.

Overall we see more signs of overfitting as we sharpen our analysis to highlight smaller effect sizes. So while we cannot rule out every form of overfitting, we view our findings as evidence that overfitting did not pose a significant danger in the most popular classification accuracy competitions on Kaggle. In spite of up to 35,000 submissions to these competitions, there are no large overfitting effects. So while overfitting may have occurred to a small extent, it did not invalidate the overall conclusions from the competitions such as which submissions rank among the top or how well they perform on the private test split.

4 Classification competitions with further evaluation metrics

In addition to classification accuracy competitions, we also surveyed competitions evaluated with AUC, MAP@K, LogLoss, and MulticlassLoss. Unfortunately, the Meta Kaggle dataset contains only *aggregate* scores for the public and private test set splits, not the loss values for individual predictions. For the accuracy metric, aggregate scores are sufficient to compute statistical measures of uncertainty such as the error bars in the scatter plots (Sections 3.1 & 3.2) or the p-value plots (Section 3.3). However, the aggregate data is insufficient to compute similar quantities for the other classification metrics. For instance, the lack of example-level scores precludes the use of standard tools such as the bootstrap or permutation tests, as we are unable to re-sample the test set. Hence our analysis here is more qualitative than for accuracy competitions in the preceding section. Nevertheless, inspecting the scatter plots can still convey overall trends in the data.

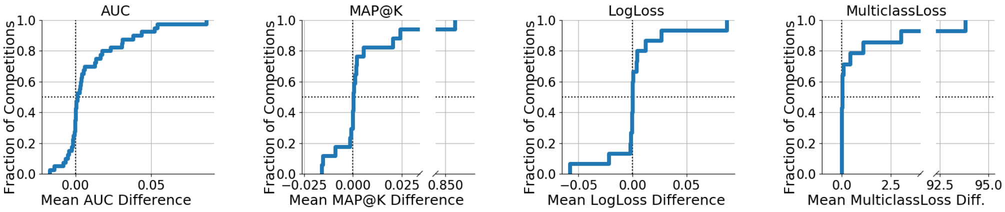


Figure 6: Empirical CDF of mean score differences (across all pre-deadline submissions) for 40 AUC competitions, 17 MAP@K competitions, 15 LogLoss competitions, and 14 MulticlassLoss competitions.

Appendices C to F contain the scatter plots for all submissions and for the top 10% of submissions to these competitions. The overall picture is similar to the accuracy plots: many competitions have scatter plots with a good linear fit close to the $y = x$ diagonal. There is more variation in the top 10% plots but due to the lack of error bars it is difficult to attribute this to overfitting vs. random noise. As before, a small number of competitions show more variation that may be indicative of overfitting. In all cases, the Kaggle website (data descriptions and discussion forums) give possible reasons for this behavior (non-i.i.d. splits, competitions with two stages and different test sets, etc.). Thus, we view these competitions as outliers that do not contradict the overall trend.

Figure 6 shows plots with aggregate statistics (mean score difference) similar to Section 3.4. As for the accuracy competitions, the empirical distribution has an approximately symmetric part centered at 0 (no score change) and a tail with larger score differences that consists mainly of outlier competitions.

5 Related work

As mentioned in the introduction, the reproducibility experiment of Recht et al. [16] also points towards a surprising absence of adaptive overfitting in popular machine learning benchmarks. However, there are two important differences to our work. First, Recht et al. [16] assembled new test sets from scratch, which makes it hard to disentangle the effects of adaptive overfitting and distribution shifts. In contrast, most of the public / private splits in Kaggle competitions are i.i.d., which removes distribution shifts as a confounder. Second, Recht et al. [16] only investigate two image classification benchmarks on which most models come from the same model class (CNNs) [9, 11]. We survey 120 competitions on which the Kaggle competitors experimented with a broad range of models and training approaches. Hence our conclusions about overfitting apply to machine learning more broadly.

The adaptive data analysis literature [6, 17] provides a range of theoretical explanations for how the common machine learning workflow may implicitly mitigate overfitting [3, 8, 12, 23]. Our work is complementary to these papers and conducts a purely empirical study of overfitting in machine learning competitions. We hope that our findings can help test and refine the theoretical understanding of overfitting in future work.

The Kaggle community has analyzed competition "shake-up", i.e., rank changes between the public and private leaderboards of a competition. We refer the reader to the comprehensive data analysis conducted by Trotman [21] as a concrete example. Focusing on rank changes is complementary to our approach based on submission scores. From a competition perspective where the winning submissions are defined by the private leaderboard rank, large rank changes are indeed undesirable. However, from the perspective of adaptive overfitting, large rank changes can be a natural consequence of random noise in the evaluation. For instance, consider a setting where a large number of competitors submits solutions with very similar public leaderboard scores. Due to the limited size of the public and private test sets, the public scores are only approximately equal to the private scores (even in the absence of any adaptive overfitting), which can lead to substantial rank changes even though all score deviations are small and of (roughly) equal size. Since the ranking approach can result in such "false positives" from the perspective of adaptive overfitting, we decided not to investigate rank changes in our paper. Nevertheless, we note that we also manually compared the shake-up results to our analysis and generally found agreement among the set of problematic competitions (e.g., competitions with non-i.i.d. splits or known public / private splits).

6 Conclusion and future work

We surveyed 120 competitions on Kaggle covering a wide range of classification tasks but found little to no signs of adaptive overfitting. Our results cast doubt on the standard narrative that adaptive overfitting is a significant danger in the common machine learning workflow.

Moreover, our findings call into question whether common practices such as limiting test set re-use increase the reliability of machine learning. We have seen multiple competitions where a non-i.i.d. split lead to substantial gaps between public and private scores, suggesting that distribution shifts [7, 15, 16, 20] may be a more pressing problem than test set re-use in current machine learning.

There are multiple directions for empirically understanding overfitting in more detail. Our analysis here focused on classification competitions, but Kaggle also hosts many regression competitions. Is there more adaptive overfitting in regression? Answering this questions will likely require access to the individual predictions of the Kaggle submissions to appropriately handle outlier submissions. In addition, there are still questions among the classification competitions. For instance, one refinement of our analysis here is to obtain statistical measures of uncertainty for competitions evaluated with metrics such as AUC (which will also require a more fine-grained version of the Kaggle data). Finally, another important question is whether other competition platforms such as CodaLab [1] or EvalAI [22] also show little signs of adaptive overfitting.

Acknowledgments

This research was generously supported in part by ONR awards N00014-17-1-2191, N00014-17-1-2401, and N00014-18-1-2833, the DARPA Assured Autonomy (FA8750-18-C-0101) and Lagrange (W911NF-16-1-0552) programs, a Siemens Futuremakers Fellowship, and an Amazon AWS AI Research Award.

References

[1] CodaLab. `https://competitions.codalab.org/competitions/`.

[2] Kaggle. `https://www.kaggle.com/`.

[3] A. Blum and M. Hardt. The Ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning (ICML)*, 2015.

[4] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[6] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Preserving statistical validity in adaptive data analysis. In *ACM symposium on Theory of computing (STOC)*, 2015.

[7] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry. Exploring the landscape of spatial robustness. In *International Conference on Machine Learning (ICML)*, 2019.

[8] V. Feldman, R. Frostig, and M. Hardt. The advantages of multiple classes for reducing overfitting from test set reuse. *International Conference on Machine Learning (ICML)*, 2019.

[9] K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. In *Biological Cybernetics*, 1980.

[10] A. Krizhevsky. Learning multiple layers of features from tiny images. `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`, 2009.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[12] H. Mania, J. Miller, L. Schmidt, M. Hardt, and B. Recht. Model similarity mitigates test set overuse. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[13] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. Chapter 1.4.8.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 2011.

[15] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

[16] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, 2019.

[17] A. Roth and A. Smith. Lectures notes “The Algorithmic Foundations of Adaptive Data Analysis”, 2017. `https://adaptivedataanalysis.com/`.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and F.-F. Li. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

[19] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[20] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[21] J. Trotman. Meta kaggle: Competition shake-up. 2019. URL `https://www.kaggle.com/jtrotman/meta-kaggle-competition-shake-up`.

[22] D. Yadav, R. Jain, H. Agrawal, P. Chattopadhyay, T. Singh, A. Jain, S. Singh, S. Lee, and D. Batra. EvalAI: Towards better evaluation systems for AI agents. 2019. URL `http://arxiv.org/abs/1902.03570`.

[23] T. Zrnic and M. Hardt. Natural analysts in adaptive data analysis. *International Conference on Machine Learning (ICML)*, 2019.

A Overview of Kaggle competitions and our resulting selection criteria

At the time of writing, the MetaKaggle dataset has 1,461 competitions. Due to the variety of different scoring mechanisms used on Kaggle and intricacies of individual competitions, it is challenging to analyze the entire datast in a coherent way. To limit the scope of our investigation and allow for a detailed analysis, we restricted our attention to a subset of competitions based on the following two criteria.

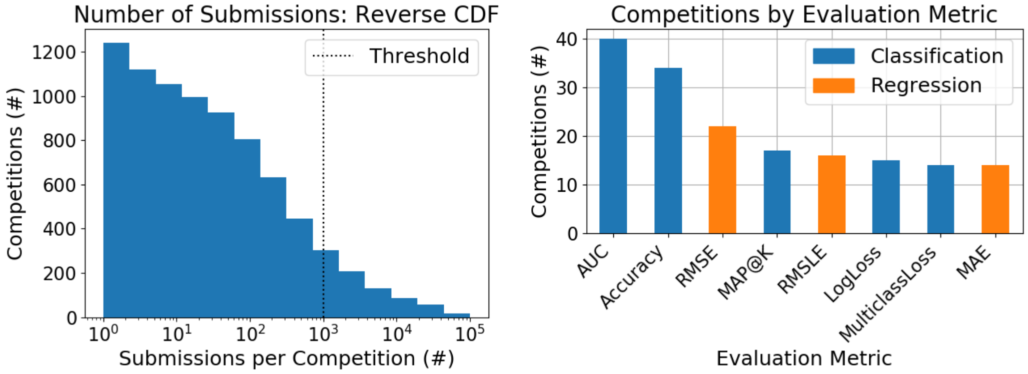


Figure 7: Overview of the Kaggle competitions. The left plot shows the distribution of submissions per competition. The right plot shows the score types that are most common among the competitions with at least 1,000 submissions.

Number of submissions. Kaggle competitions have a wide range of submissions, ranging from 0 for the bottom 10% to 3,159 for the most popular 10% of competitions (and 132,097 for the most popular competition). The left plot in Figure 7 shows the distribution of competition sizes. From the perspective of adaptive overfitting, competitions with many submissions are more interesting since a larger number of submissions indicate more potential adaptivity. Moreover, overfitting on the most popular competitions would also be more important to understand as it would undermine the credibility of the competition paradigm. So to limit the number of competitions, we included only competitions with at least 1,000 submissions before the corresponding competition deadline. This left a total of 262 competitions.

Score type. Depending on the competition, the performance of a submissions is measured with one of various scoring mechanisms. The list of scoring mechanisms involves well-known score types (accuracy, AUC, RMSE) and a long tail of more specialized metrics (see the right plot in Figure 7. We focus on score types used for classification tasks with at least 10 competitions (passing our submission count filter) so we can compare competitions with multiple other competitions using the same score type. This avoids difficulties such as comparing classification vs. regression competitions or bounded vs. unbounded scores. Moreover, the most popular score types are also the most relevant for measuring the amount of overfitting in machine learning competitions. The classification score types with at least 10 competitions are classification accuracy, AUC (area under curve), MAP@K (mean average precision), the logistic loss, and a multiclass variant of the logistic loss.

In total, the above exclusion rules leave 120 competitions. It would be possible to restrict the set of competitions even further, e.g., by only including "featured" competitions (the competitions promoted by Kaggle, usually with prize money) and excluding other competition types such as "in class" competitions which are mainly used for educational purposes. Another possible filter would be to exclude competitions with only small test set sizes since the resulting scores are inherently more noisy. However, we decided to minimize the exclusion rules so we can analyze a large number of competitions and possibly find effects of features such as competition type and test set size.

B Accuracy

B.1 Accuracy: competition info

Table 2: Competitions scored with accuracy with greater than 1000 submissions. n_{public} is the size of the public test set and n_{private} is the size of the private test set. N/A means that we could not access the competition data to compute the dataset sizes. A * after the competition name means the name was slightly edited to fit into the table.

Accuracy				
ID	Name	# Sub.	n_{public}	n_{private}
3362	Dogs vs. Cats	1,225	3,750	8,751
3366	The Black Box Learning Challenge*	1,924	5,000	5,000
3428	Data Science London + Scikit-learn	1,477	2,700	6,299
3641	DecMeg2014 - Decoding the Human Brain	4,507	1,745	2,313
3649	CIFAR-10 - Object Recognition in Images	1,634	300	9,700
3788	Allstate Purchase Prediction Challenge	24,532	59,657	139,199
4554	MSU Visits*	1,288	30,000	270,000
4821	104-1 MLDS_Final	1,060	36,400	36,401
4857	Let's Overfit	1,080	500,000	500,000
5275	Can we predict voting outcomes?	35,247	696	696
5896	Prediction Reviews Sentiment Analysis*	1,770	50	50
5903	Prediction Reviews Sentiment Analysis (Light)*	3,528	400	100
5942	Identify Me If You Can – Yandex & MIPT	1,242	20,588	20,589
6109	Identify Me If You Can	1,819	23,236	23,237
7042	Text Normalization Challenge - English Language	1,834	20,446	2,024,166
7115	Cdiscount's Image Classification Challenge	5,859	530,455	1,237,727
7563	Cover Type Prediction of Forests	3,077	6,000	14,000
7634	TensorFlow Speech Recognition Challenge	24,263	3,171	155,365
8307	Data Science Nigeria Telecoms Churn	1,259	180	420
9051	Digit Recongnizer 1438	1,640	2,576	2,575
9491	ML-2018spring-hw5	2,247	100,000	100,000
10851	CSE158 fa18 Category Prediction	3,872	7,000	7,000
11842	AIA image classification by CNN	3,415	1,350	150
12187	CS 4740 Project 4 Revised	1,086	N/A	N/A
12349	2019 1st ML month with KaKR	3,307	209	209
12357	UC Berkeley CS189 HW1 (MNIST)	1,480	3,000	7,000
12358	UC Berkeley CS189 HW1 (SPAM)	1,527	1,757	4,100
12359	UC Berkeley CS189 HW1 (CIFAR-10)	1,336	3,000	7,000
12598	TAU Robot Surface Detection	1,574	852	853
12681	PadhAI: Text - Non Text Classification Level 2	1,185	90	210
13034	UC Berkeley CS189 HW3 (MNIST)	1,193	3,000	7,000
13036	UC Berkeley CS189 HW3 (SPAM)	1,080	1,757	4,100
13078	Homework 2 Part 2 - Classification	2,245	N/A	N/A
13383	ML2019spring-hw2	2,802	8,141	8,141

B.2 Accuracy: outlier competitions

Figure 8 shows the plots corresponding to the three levels of overfitting analysis for the mean accuracy difference outliers discussed in Section 3.4.

By reading the Kaggle forums and the description of the dataset construction, we were able to determine a possible cause of overfitting for most of the outlier competitions. In some cases, the competition hosts did not create the public and private test splits in a truly i.i.d. manner, and competitors were able to exploit the difference in distribution. In other cases, the size of the test set was extremely small. The list of outlier competitions and corresponding cause of overfitting follow. We also include links to relevant descriptions of the data or discussions on the Kaggle forums.

- **Competition 3641: DecMeg2014 - Decoding the Human Brain** The test set contains data generated from 7 subjects and the public and private splits were created in a non-i.i.d. manner using data from individual subjects. `https://www.kaggle.com/c/decoding-the-human-brain/data`
- **Competition 5903: Prediction Reviews Sentiment Analysis (Light)** The competition uses a small test set consisting of 500 examples (400 in the public test set and 100 in the private test set).
- **Competition 12349: 2019 1st ML month with KaKR** The competition uses a small test set consisting of 418 examples (209 in the private test set and 209 in the public test set.)
- **Competition 12598: TAU Robot Surface Detection** The dataset is a time series that consists of measurements of acceleration, velocity and orientation. In some of the measurement locations, the floor has a slope that makes the orientation channels informative (though the orientation terms are designed to not be informative for the particular task). In order to mitigate this effect, the organizers split the time series data into shorter subsequences and assigned separate subsequences to the public and private split. Thus, the public and private data were not created from a truly i.i.d. split. [4]
- **Competition 12681: PadhAI: Text - Non Text Classification Level 2** The competition uses a small test set consisting of 300 examples (90 in the public test set and 210 in the private test set).

---

[4] Personal communication with the competition organizer Heikki Huttunen.

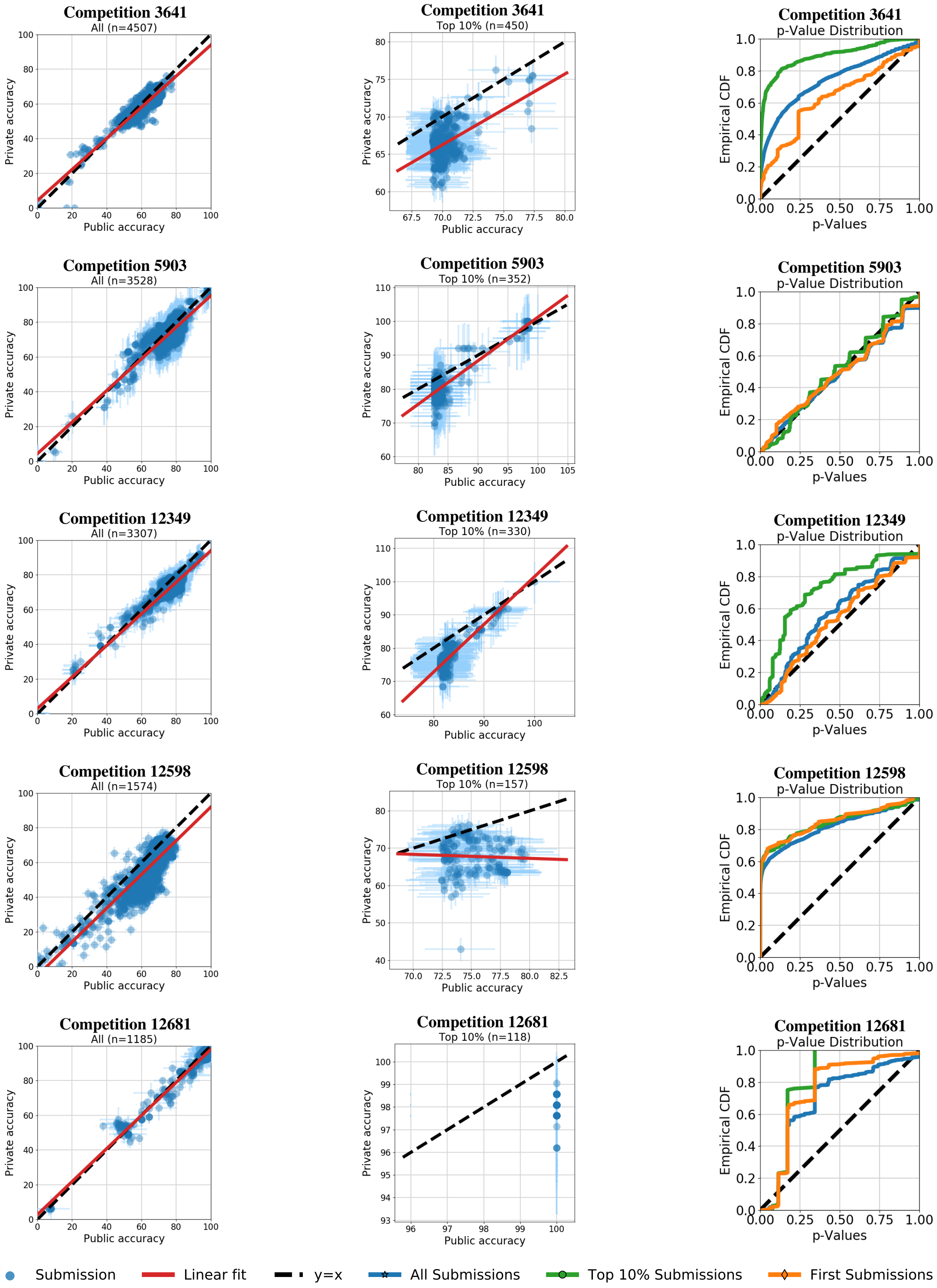


Figure 8: The three main plots for the outliers in mean score difference discussed in Section 3.4: private versus public accuracy for all submissions (left), the top 10% of submissions ranked by public accuracy (center), and the empirical CDFs of p-Values (right). For the scatter plots, each point corresponds to an individual submission (shown with 95% Clopper-Pearson confidence intervals). The right figure includes empirical CDFs for an idealized null model that assumes no overfitting (see Section 3.3).

B.3 Accuracy: all submissions

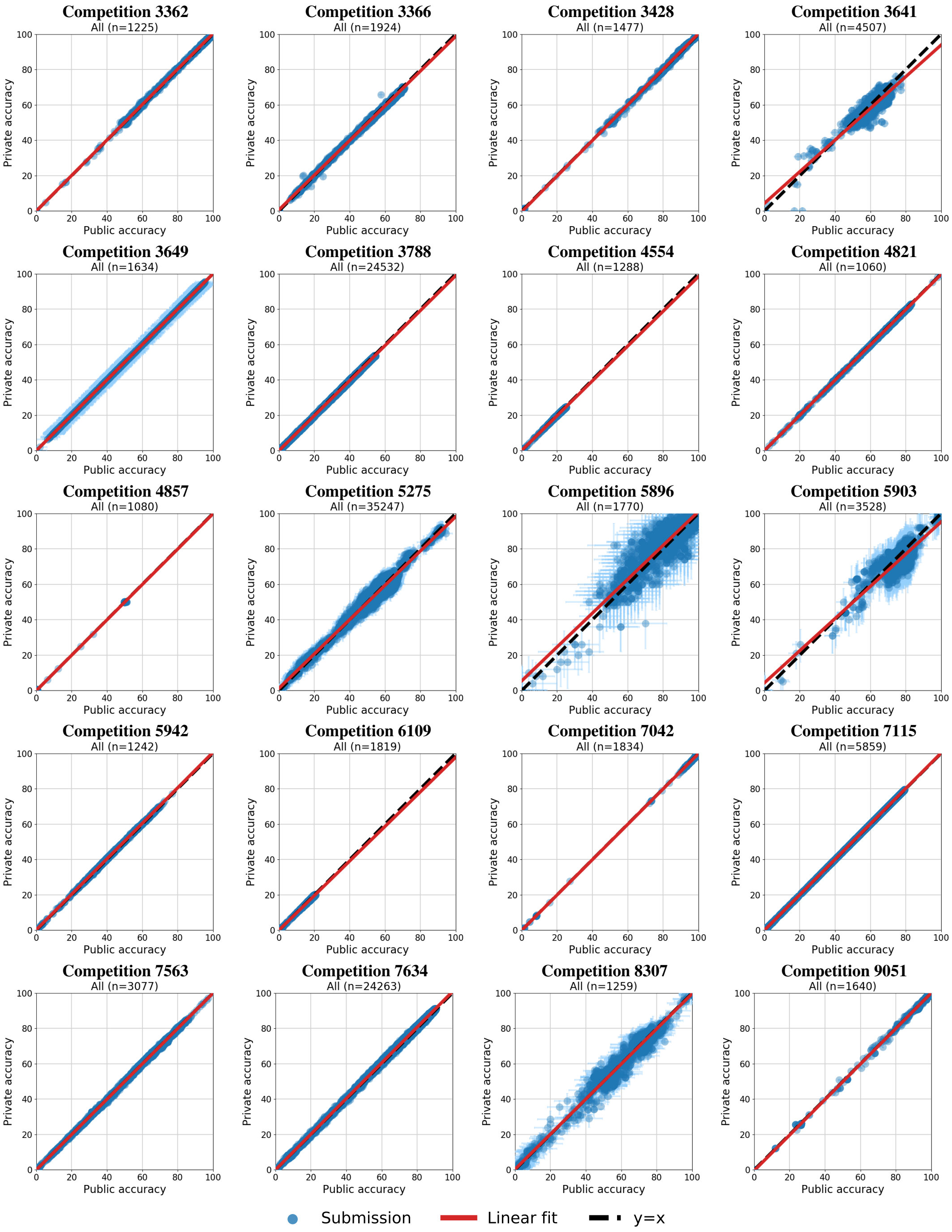

Competition 3362
All (n=1225)
Competition 3366
All (n=1924)
Competition 3428
All (n=1477)
Competition 3641
All (n=4507)
Competition 3649
All (n=1634)
Competition 3788
All (n=24532)
Competition 4554
All (n=1288)
Competition 4821
All (n=1060)
Competition 4857
All (n=1080)
Competition 5275
All (n=35247)
Competition 5896
All (n=1770)
Competition 5903
All (n=3528)
Competition 5942
All (n=1242)
Competition 6109
All (n=1819)
Competition 7042
All (n=1834)
Competition 7115
All (n=5859)
Competition 7563
All (n=3077)
Competition 7634
All (n=24263)
Competition 8307
All (n=1259)
Competition 9051
All (n=1640)

Figure 9: Private versus public accuracy for all submissions for the most popular Kaggle accuracy competitions. Each point corresponds to an individual submission (shown with 95% Clopper-Pearson confidence intervals).

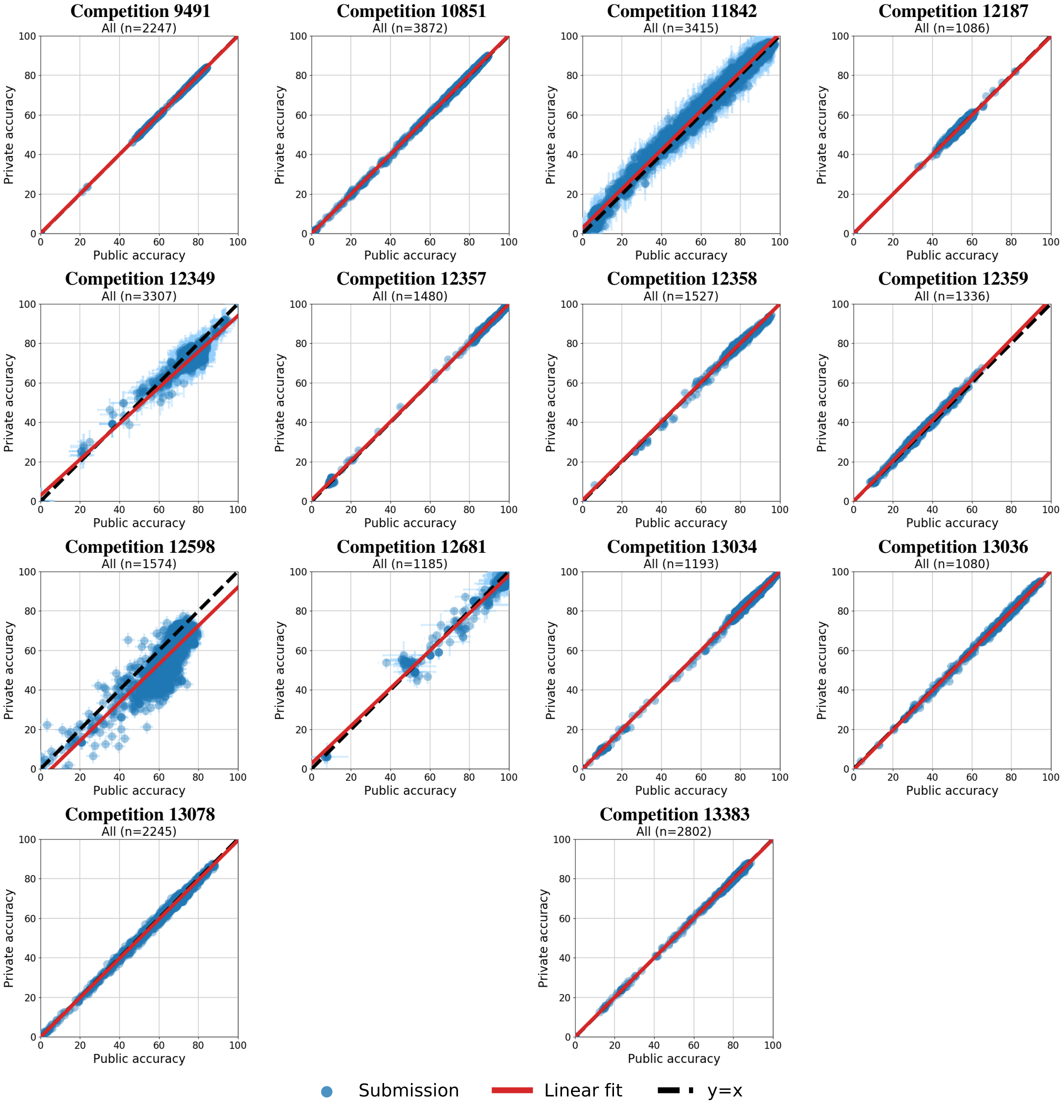


Figure 9: Private versus public accuracy for all submissions for the most popular Kaggle accuracy competitions. Each point corresponds to an individual submission (shown with 95% Clopper-Pearson confidence intervals).

B.4 Accuracy: top 10% of submissions

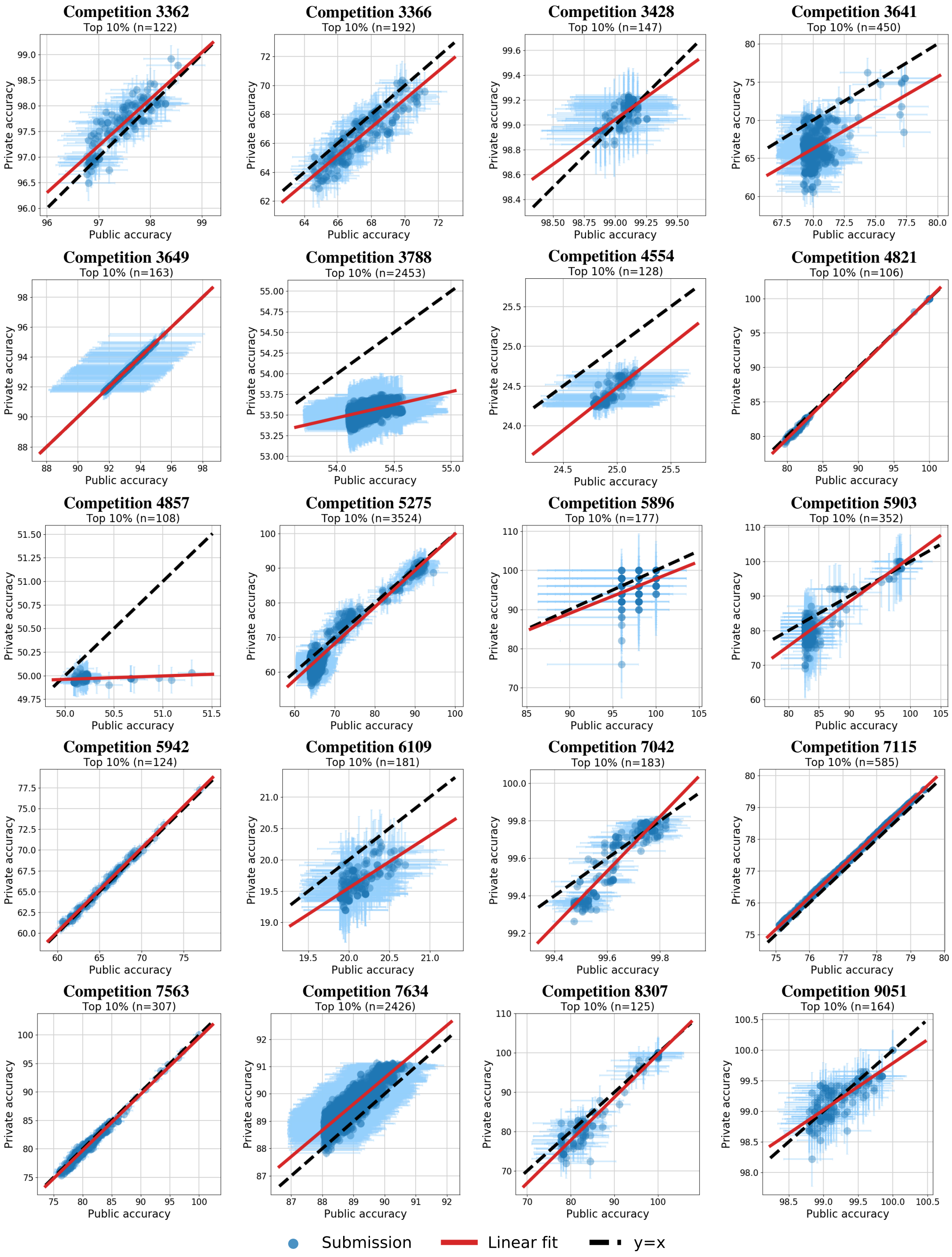


Figure 10: Private versus public accuracy for top 10% of submissions for the most popular Kaggle accuracy competitions. Each point corresponds to an individual submission (shown with 95% Clopper-Pearson confidence intervals).

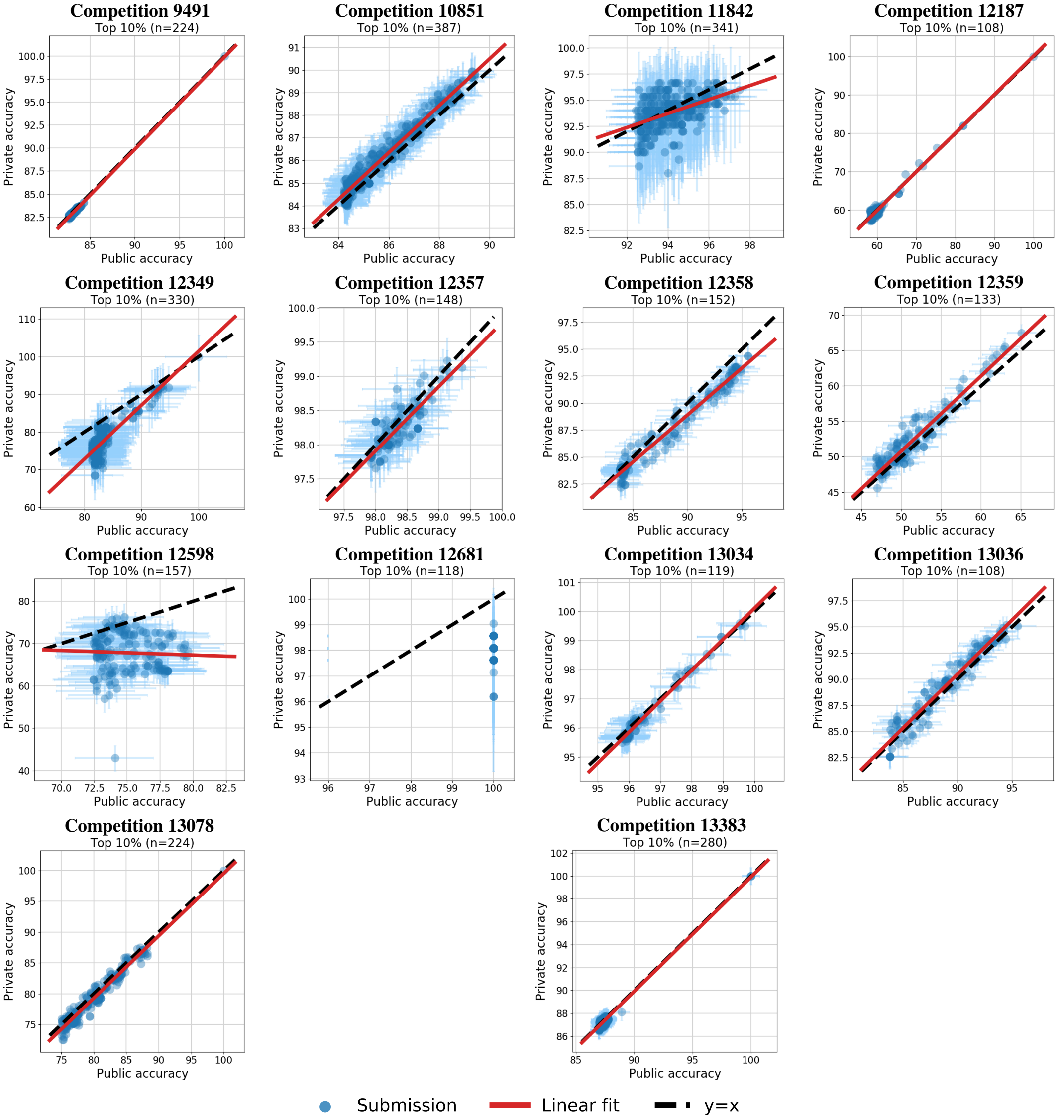


Figure 10: Private versus public accuracy for all submissions for the most popular Kaggle accuracy competitions. Each point corresponds to an individual submission (shown with 95% Clopper-Pearson confidence intervals).

B.5 Accuracy: p-values

B.5.1 Computation of P-Values

For a given submission, the public and private test accuracies will rarely exactly match. We would like to understand exactly how much of this discrepancy can be attributed to random variation as opposed to overfitting. Using only MetaKaggle data, we compute *exact* p-values for a simple null mode.

Fix a classifier that correctly classifies m points out of a total test set of size n_{total} and incorrectly classifies the remaining $n_{\text{total}} - m$ points. We randomly split these n_{total} points into public and private test sets of size n_{public} and n_{private}, respectively, with $n_{\text{public}} + n_{\text{private}} = n_{\text{total}}$. Under the null model, every partition of the points is equally likely. Dividing points in this fashion leads to m_{public} and m_{private} successes on each subset. Let Δ be the observed difference in public and private accuracies. Then, we're interested in computing

$$\texttt{p\_value} = \Pr\left\{\left|\frac{m_{\text{private}}}{n_{\text{private}}} - \frac{m_{\text{public}}}{n_{\text{public}}}\right| > \Delta\right\}. \tag{3}$$

Fixing m_{private} also determines m_{public}, and for any fixed $m_{\text{private}} = k$, the probability of sampling a private test set of with k correctly classified points is

$$\frac{\binom{m}{k}\binom{n-m}{n_{\text{private}}-k}}{\binom{n}{n_{\text{private}}}}. \tag{4}$$

Therefore, we can compute the p-value in equation (3) as

$$\Pr\left\{\left|\frac{m_{\text{private}}}{n_{\text{private}}} - \frac{m_{\text{public}}}{n_{\text{public}}}\right| > \Delta\right\} = \sum_{k \in [m] : \left|\frac{k}{n_{\text{private}}} - \frac{(m-k)}{n_{\text{public}}}\right| > \Delta} \frac{\binom{m}{k}\binom{n-m}{n_{\text{private}}-k}}{\binom{n}{n_{\text{private}}}}, \tag{5}$$

where $[m] = \{0, 1, 2, \ldots, m\}$. Evaluating this summation only requires access to $n_{\text{total}}, n_{\text{public}}, n_{\text{private}}$, and Δ, all of which are available in the MetaKaggle dataset. Hence, a p-value can be readily evaluated for each submission. Computing ratios of binomial coefficients like (4) is numerically unstable and computationally expensive for large n. Therefore, in our implementation, we work in log-space and use Stirling's approximation for the binomial coefficients to efficiently and reliably approximate (5).

B.5.2 Accuracy: p-value plots

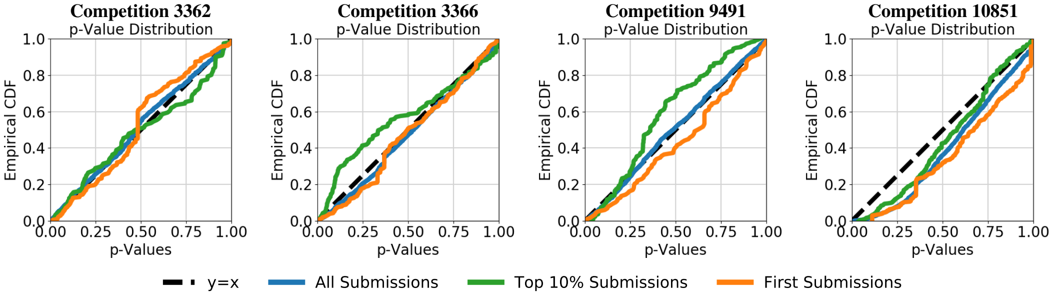


Figure 11: Empirical CDFs for an idealized null model that assumes no overfitting (full details in Section 3.3).

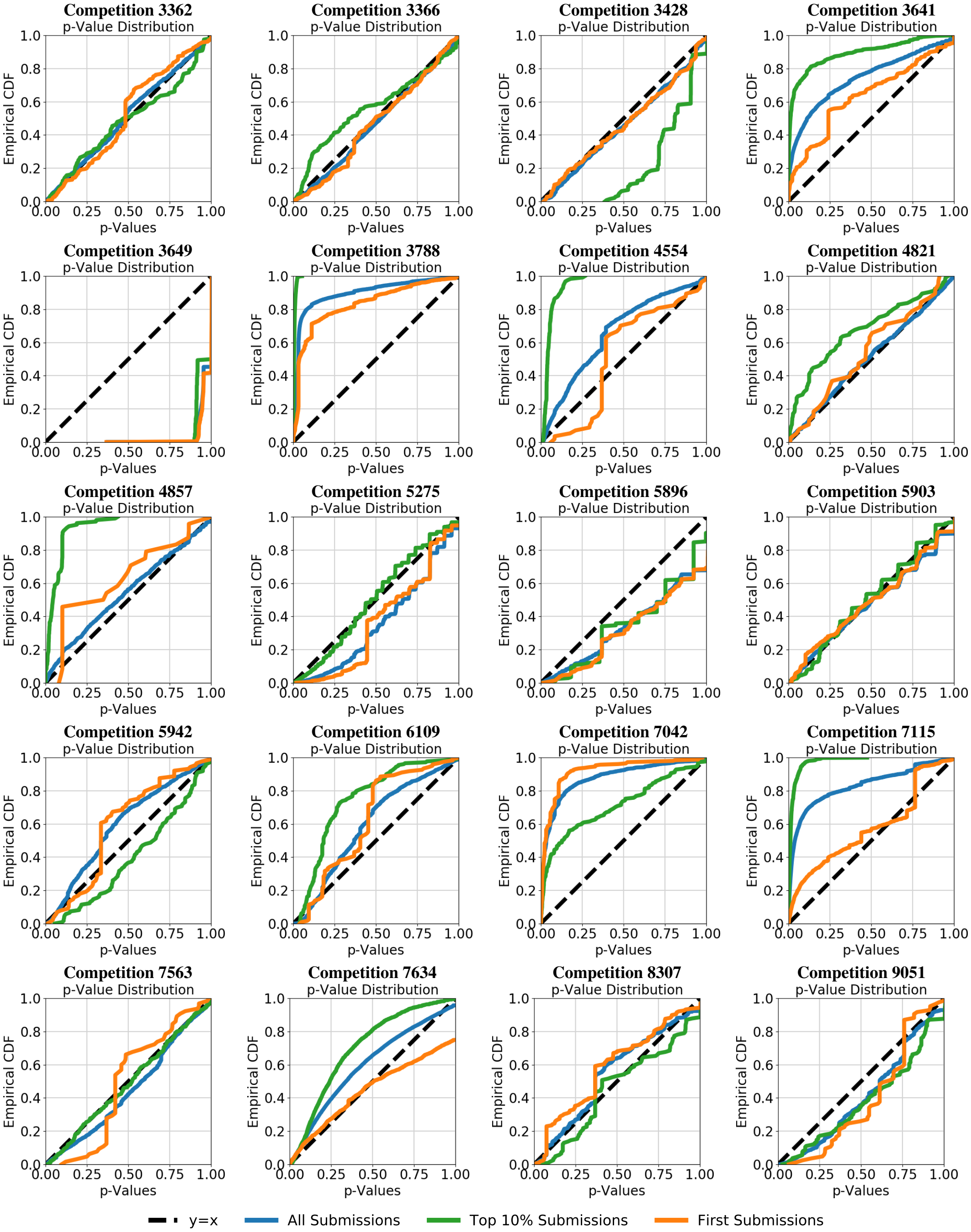


Figure 12: Empirical CDFs for an idealized null model that assumes no overfitting (full details in Section 3.3)

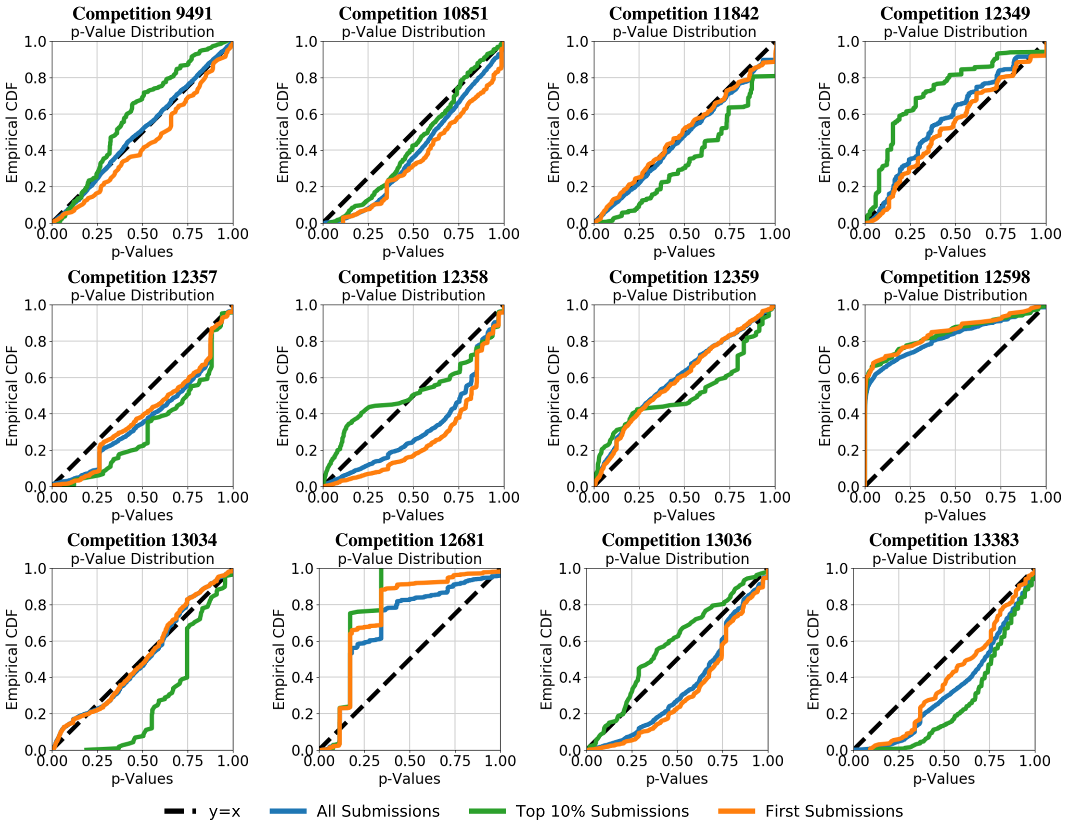


Figure 12: Empirical CDFs for an idealized null model that assumes no overfitting (full details in Section 3.3)

C AUC

C.1 AUC: competition info

Table 3: Competitions scored with AUC with greater than 1000 submissions. n_{public} is the size of the public test set and n_{private} is the size of the private test set. N/A means that we could not access the competition data to compute the dataset sizes. A * after the competition name means the name was slightly edited to fit into the table.

AUC				
ID	Name	# Sub.	n_{public}	n_{private}
2439	INFORMS Data Mining Contest 2010	1,483	254	2,285
2445	Predict Grant Applications	2,800	544	1,632
2464	IJCNN Social Network Challenge	1,124	1,792	7,168
2478	Stay Alert! The Ford Challenge	1,402	36,252	84,588
2489	Don’t Overfit!	3,775	1,975	17,775
2551	Give Me Some Credit	7,724	30,451	71,052
3338	Amazon.com - Employee Access Challenge	16,896	17,676	41,245
3353	The Marinexplore and Cornell Univ. Whale Detection*	3,295	16,351	38,152
3469	Influencers in Social Networks	2,109	2,976	2,976
3509	The ICML 2013 Whale Challenge - Right Whale Redux	1,005	7,641	17,828
3524	Accelerometer Biometric Competition	7,130	8,102,160	18,905,040
3526	StumbleUpon Evergreen Classification Challenge	7,509	634	2,537
3774	CONNECTOMICS	1,458	75,742	75,743
3897	Acquire Valued Shoppers Challenge	25,205	N/A	N/A
3926	Predicting Excitement at DonorsChoose.org*	12,530	11,193	33,579
3933	MLSP 2014 Schizophrenia Classification Challenge	2,246	0	119,749
3960	American Epilepsy Society Seizure Prediction Challenge	17,782	1,574	2,361
4031	Driver Telematics Analysis	36,072	N/A	N/A
4043	BCI Challenge @ NER 2015	4,348	680	2,721
4294	Facebook Recruiting IV: Human or Robot?	13,559	1,410	3,290
4366	West Nile Virus Prediction	29,963	2,326	113,967
4487	Springleaf Marketing Response	39,439	43,570	101,662
4493	Truly Native?	3,224	1,360	2,040
4657	Homesite Quote Conversion	36,387	52,151	121,685
4986	Santander Customer Satisfaction	93,584	37,909	37,909
5167	PRED 411-2016_04-U2-INSURANCE-A*	1,436	1,070	1,071
5174	Avito Duplicate Ads Detection	8,157	657,602	657,603
5261	Predicting Red Hat Business Value	33,696	149,606	349,081
5390	Melbourne Univ. Seizure Prediction*	10,083	N/A	N/A
6242	Catch Me If You Can: Intruder Detection*	2,479	41,398	41,399
7162	WSDM - KKBox’s Music Recommendation Challenge	15,555	1,278,396	1,278,395
8227	2018 Spring CSE6250 HW1	1,170	244,173	244,173
8540	TalkingData AdTracking Fraud Detection Challenge	68,594	3,382,284	15,408,185
9120	Home Credit Default Risk	132,097	9,749	38,995
10683	Microsoft Malware Prediction	43,702	4,947,549	2,905,704
11803	BGU - Machine Learning	2,079	6,849	5,603
11848	Histopathologic Cancer Detection	20,352	28,154	29,304
12512	2019 Spring CSE6250 BDH	1,396	244,173	244,173
12558	WiDS Datathon 2019	3,021	4,312	2,222
12904	Caltech CS 155 2019 Part 1	1,078	1,264,122	1,264,122

C.2 AUC: outlier competitions

For AUC, we also investigated competitions whose linear fit was a bad approximation to the diagonal $y = x$ line. We first identified these competitions using visual inspection of the plots in Section C.3.

By reading the Kaggle forums and the description of the dataset construction, we were able to determine a possible cause of overfitting for most of the outlier competitions. In most cases, the competition hosts did not create the public and private test splits in a truly i.i.d. manner, and competitiors were able to exploit the difference in distribution. In other cases, the size of the test set was extremely small. The list of outlier competitions and corresponding cause of overfitting follow. We also include links to relevant descriptions of the data or discussions on the Kaggle forums.

- **Competition 3774: CONNECTOMICS** Participants knew which test data belonged in the public and private splits because the splits were provided in separate files. `https://www.kaggle.com/c/connectomics/data`
- **Competition 3926: KDD Cup 2014 - Predicting Excitement at DonorsChoose.org**: Public and private splits are not i.i.d. and likely were created by separating donations projects by the date they were proposed. `https://www.kaggle.com/c/kdd-cup-2014-predicting-excitement-at-donors-choose/discussion/9772#latest-50827`
- **Competition 3933: MLSP 2014 Schizophrenia Classification Challenge** The test set contains data generated from 58 subjects and the public and private splits were created in a non-i.i.d. manner using data from individual subjects. `https://www.kaggle.com/c/mlsp-2014-mri/discussion/10135#latest-54483`
- **Competition 4043: BCI Challenge @ NER 2015**: The test set contains data generated from 10 subjects and the public and private splits were created in a non-i.i.d. manner using data from individual subjects `https://www.kaggle.com/c/inria-bci-challenge/discussion/12613#latest-65652`.
- **Competition 8540: TalkingData AdTracking Fraud Detection Challenge** A second, larger test set that was unintentionally released at the start of the competition and then participants were permitted but not required to use the data. `https://www.kaggle.com/c/talkingdata-adtracking-fraud-detection/discussion/52658#latest-315882`
- **Competition 10683: Microsoft Malware Prediction** The private test data included several severe outliers not present in the public test data, indicating that the public and private split are not i.i.d. `https://www.kaggle.com/c/microsoft-malware-prediction/discussion/84745`

C.3 AUC: all submissions

Figure 13 shows the private versus public AUC scatter plots for each AUC competition with over 1000 submissions in the MetaKaggle dataset.

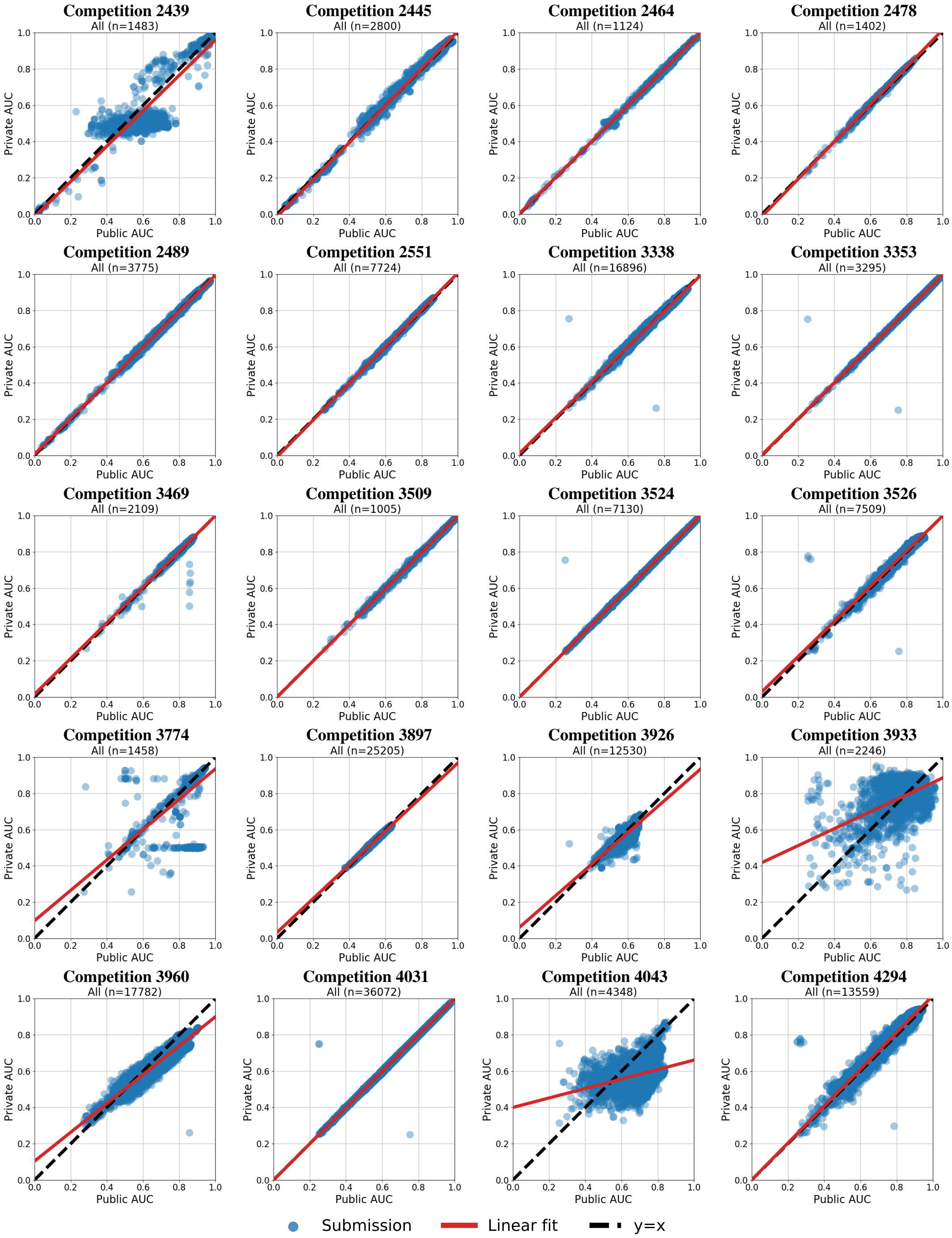


Figure 13: Private versus public AUC for all submissions for Kaggle AUC competitions.

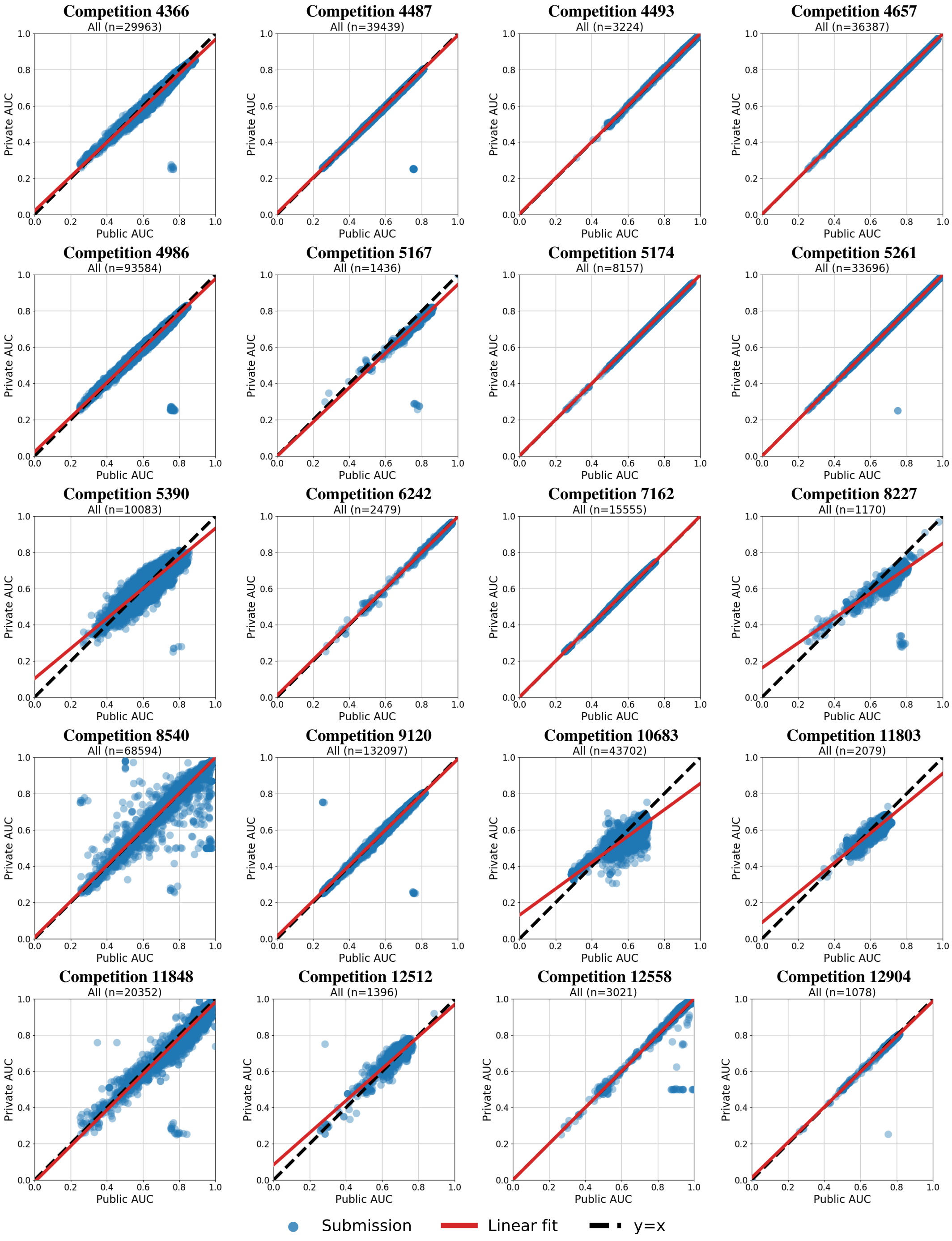
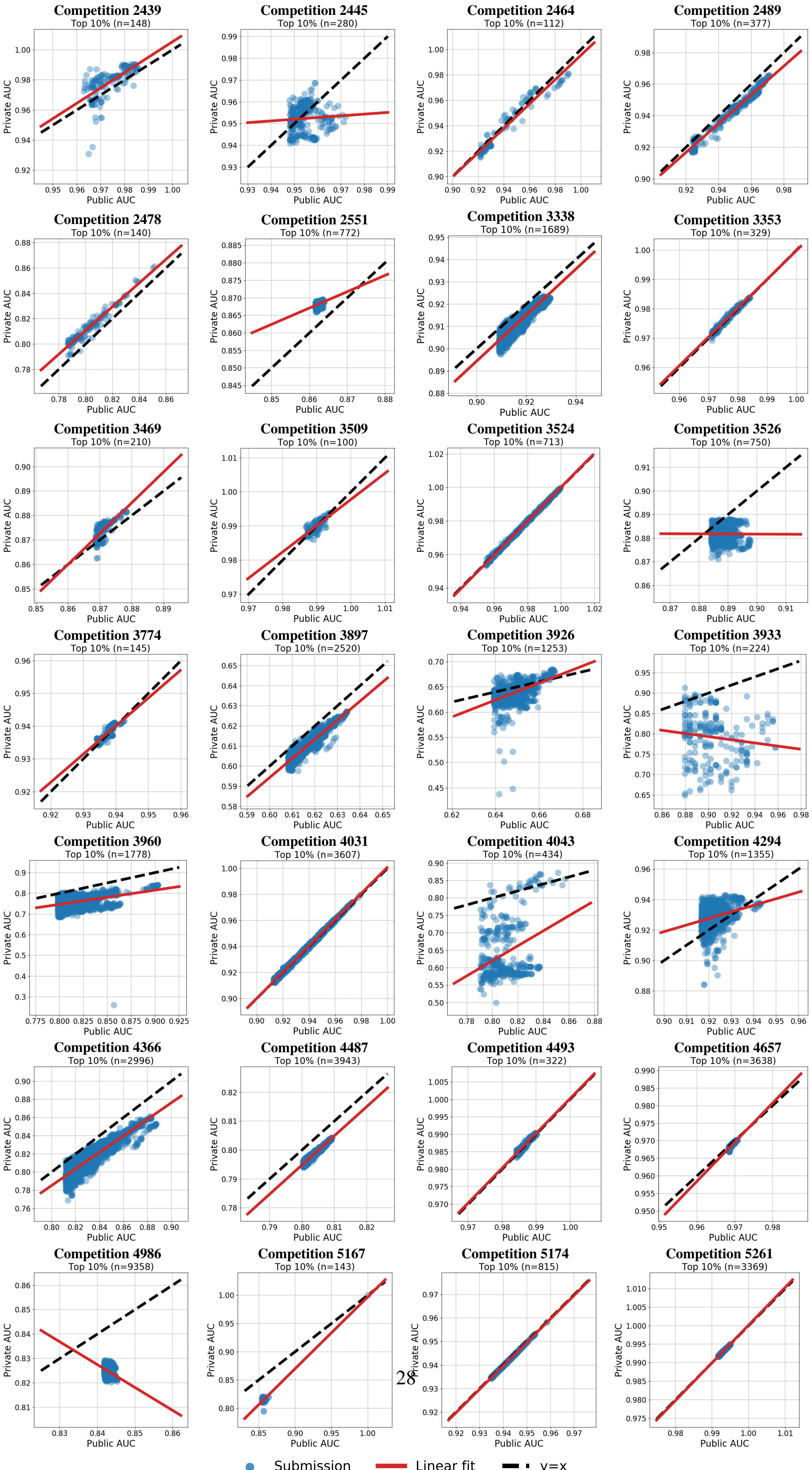


Figure 13: Private versus public AUC for all submissions for Kaggle AUC competitions.

C.4 AUC: top 10% of submissions


Competition 2439
Top 10% (n=148)
Competition 2445
Top 10% (n=280)
Competition 2464
Top 10% (n=112)
Competition 2489
Top 10% (n=377)
Competition 2478
Top 10% (n=140)
Competition 2551
Top 10% (n=772)
Competition 3338
Top 10% (n=1689)
Competition 3353
Top 10% (n=329)
Competition 3469
Top 10% (n=210)
Competition 3509
Top 10% (n=100)
Competition 3524
Top 10% (n=713)
Competition 3526
Top 10% (n=750)
Competition 3774
Top 10% (n=145)
Competition 3897
Top 10% (n=2520)
Competition 3926
Top 10% (n=1253)
Competition 3933
Top 10% (n=224)
Competition 3960
Top 10% (n=1778)
Competition 4031
Top 10% (n=3607)
Competition 4043
Top 10% (n=434)
Competition 4294
Top 10% (n=1355)
Competition 4366
Top 10% (n=2996)
Competition 4487
Top 10% (n=3943)
Competition 4493
Top 10% (n=322)
Competition 4657
Top 10% (n=3638)
Competition 4986
Top 10% (n=9358)
Competition 5167
Top 10% (n=143)
Competition 5174
Top 10% (n=815)
Competition 5261
Top 10% (n=3369)
Public AUC
Private AUC
Submission
Linear fit
y=x

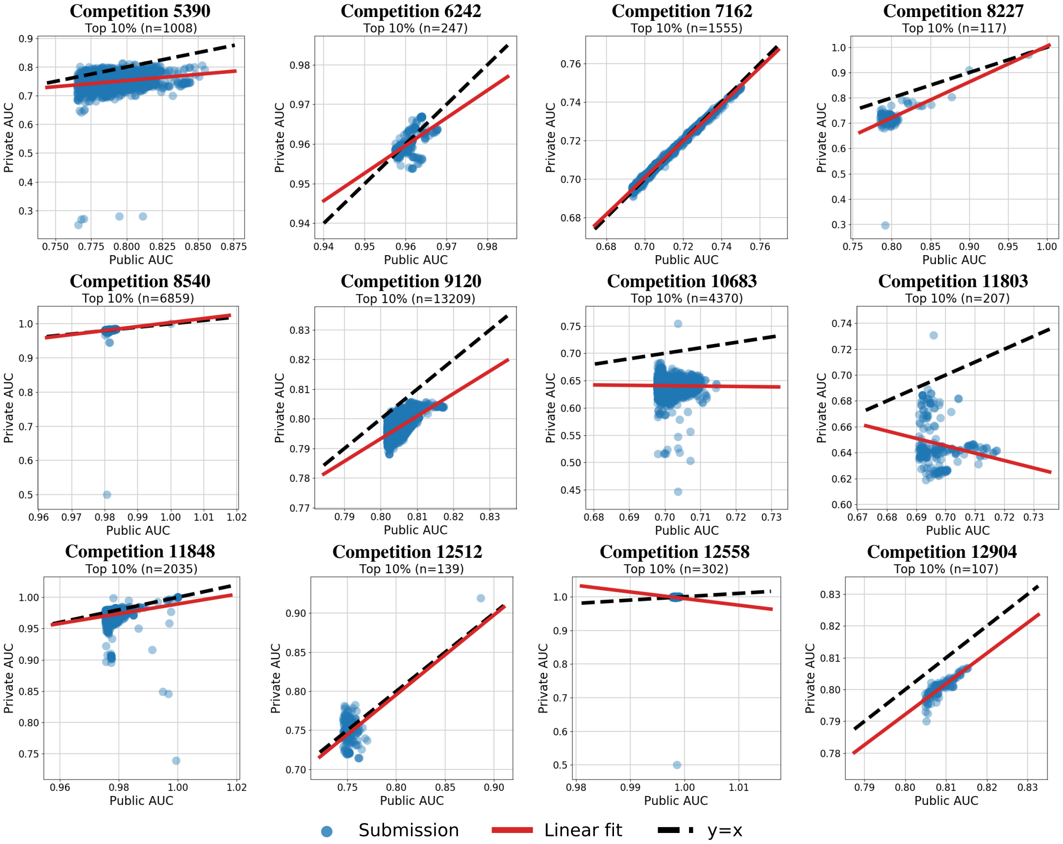


Figure 14: Private versus public AUC for top 10% of submissions for Kaggle AUC competitions.

D MAP@K

D.1 MAP@K: competition info

Our MAP@K (Mean Average Precision @ K) category includes competitions marked in the MetaKaggle dataset as MAP@{K}, MAP@k, MAP@3, AP@{K}, and MAP@{K}_OLD, as these metrics are functionally equivalent. MAP@K is typically used for tasks that require predicting a ranked list of items; for instance, listing K advertisements in order based on how likely a particular user is to click on them. In this example, the 'Mean' in 'Mean Average Precision' is taken over all users. For the i'th user, 'Average Precision' of the K predicted advertisements is defined as:

$$\frac{1}{\min(m_i, K)} \sum_{k=1}^{K} P_i(k) I_i(k),$$

where m_i is the number of advertisements that user i actually clicked on, $P_i(k)$ is the precision of the first k recommendations for user i (the fraction of the first k recommended advertisements that the user actually clicked on), and $I_i(k)$ is an indicator variable representing whether the k'th advertisement was clicked on. In the MetaKaggle dataset, MAP@{K} and MAP@k denote precisely this Mean Average Precision @ K scoring metric, for varying values of K. MAP@3 is the same, with $K = 3$. AP@{K} denotes Average Precision @ K, which is functionally equivalent to MAP@K because such competitions require predicting a single list (and the mean of a single element is itself). MAP@{K}_OLD is also effectively the same metric, except with no pre-specified limit on K. In our earlier analogy, this would correspond to recommending all the available advertisements (rather than only the top K), so the scoring metric is based solely on the ordering of the recommendations.

Table 4: Competitions scored with MAP@K with greater than 1000 submissions. n_{public} is the size of the public test set and n_{private} is the size of the private test set. A * after the competition name means the name was slightly edited to fit into the table. A ** indicates competitions where the public test set and the private test set are identical and both use all of the test data.

MAP@K				
ID	Name	# Sub.	n_{public}	n_{private}
2748	KDD Cup 2012, Track 1	13,077	710,267	629,860
2947	Facebook Recruiting Competition	3,554	65,647	196,941
3288	Event Recommendation Engine Challenge	3,021	10,238	10,238**
3445	KDD Cup 2013 (Track 1)*	9,426	2,244	2,244**
3929	The Hunt for Prohibited Content	5,001	675,622	675,621
4481	Coupon Purchase Prediction	18,487	6,870	16,030
5056	Expedia Hotel Recommendations	22,713	834,321	1,693,923
5186	Facebook V: Predicting Check Ins	15,127	4,445,370	4,445,369
5497	Outbrain Click Prediction	6,654	9,667,549	22,557,613
5558	Santander Product Recommendation	28,773	278,884	650,731
6818	Humpback Whale Identification	37,529	1,592	6,368
7666	CS5785 Fall 2017 Final Exam	3,898	1,000	1,000
8396	Google Landmark Retrieval Challenge	3,123	40,019	77,685
8857	Recipient prediction 2018	1,154	1,000	1,000
8900	Freesound Audio Tagging Challenge*	5,684	282	9,118
10200	Quick, Draw! Doodle Recognition Challenge	21,407	10,098	102,101
12088	CS5785 Fall 2018 Final	1,877	1,000	1,000

D.2 MAP@K: outlier competitions

For MAP@K, we also investigated competitions whose linear fit was a bad approximation to the diagonal $y = x$ line. We first identified these competitions using visual inspection of the plots in Section D.3.

By reading the Kaggle forums and the description of the dataset construction, we were able to determine a possible cause of overfitting for most of the outlier competitions. In most cases, the public and private test set splits were known to the participants. We also include links to relevant descriptions of the data or discussions on the Kaggle forums.

- **Competition 2748: KDD Cup 2012, Track 1** Participants knew which test data belonged in the public and private splits because the splits were based on timestamp. `https://www.kaggle.com/c/kddcup2012-track1/discussion/1632#latest-9759`
- **Competition 3288: Event Recommendation Engine Challenge** The public and private test sets were released in separate stages, so in the final stage of the competition participants knew which test data belonged in each split. `https://www.kaggle.com/c/event-recommendation-engine-challenge/discussion/3840#latest-20556`
- **Competition 3445: KDD Cup 2013 - Author-Paper Identification Challenge (Track 1)** Participants knew which test data belonged in the public and private splits because the splits were released at different stages of the competition. `https://www.kaggle.com/c/kdd-cup-2013-author-paper-identification-challenge/overview/timeline`

D.3 MAP@K: all submissions

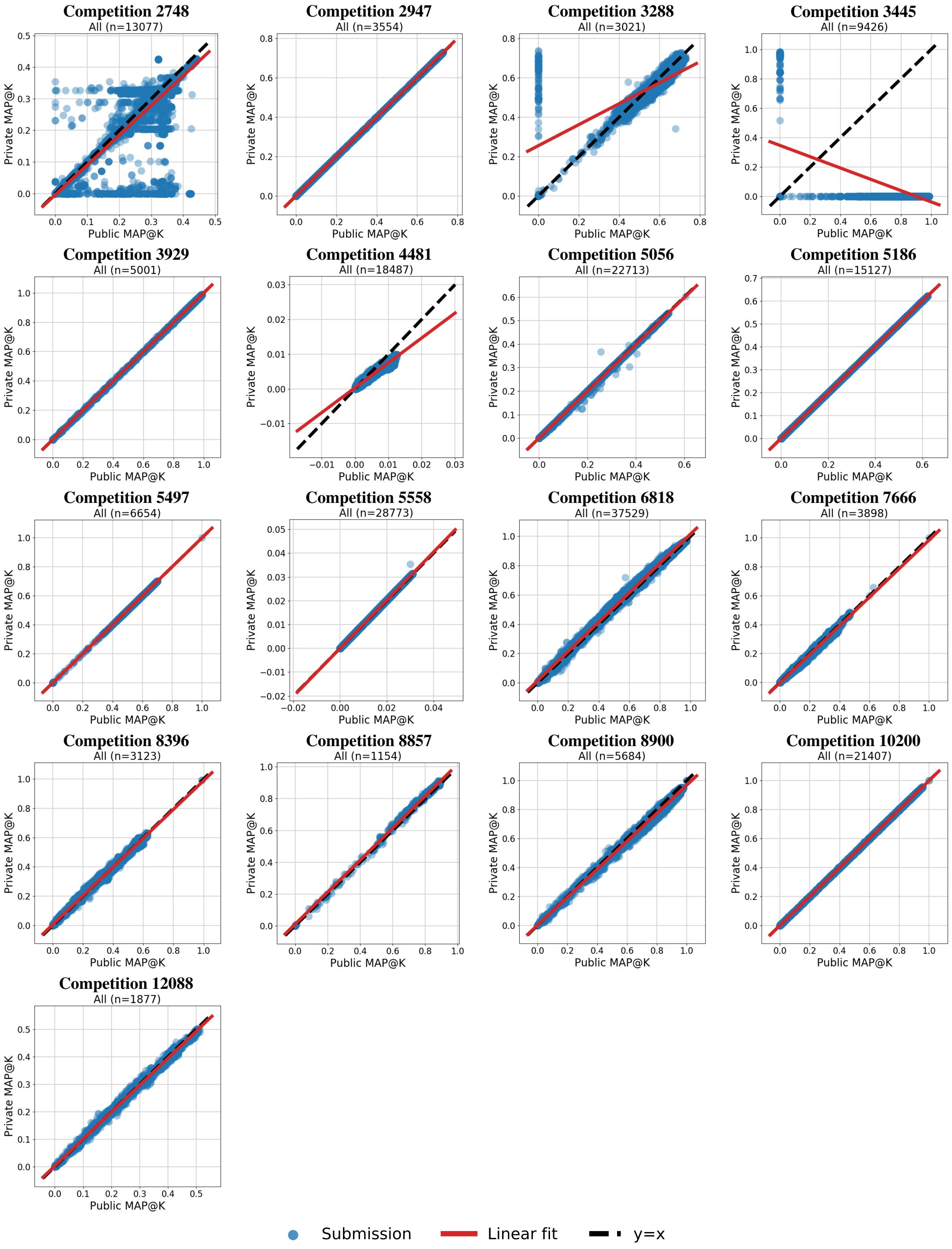


Figure 15: Private versus public MAP@K for all submissions for Kaggle MAP@K competitions.

D.4 MAP@K: top 10% of submissions

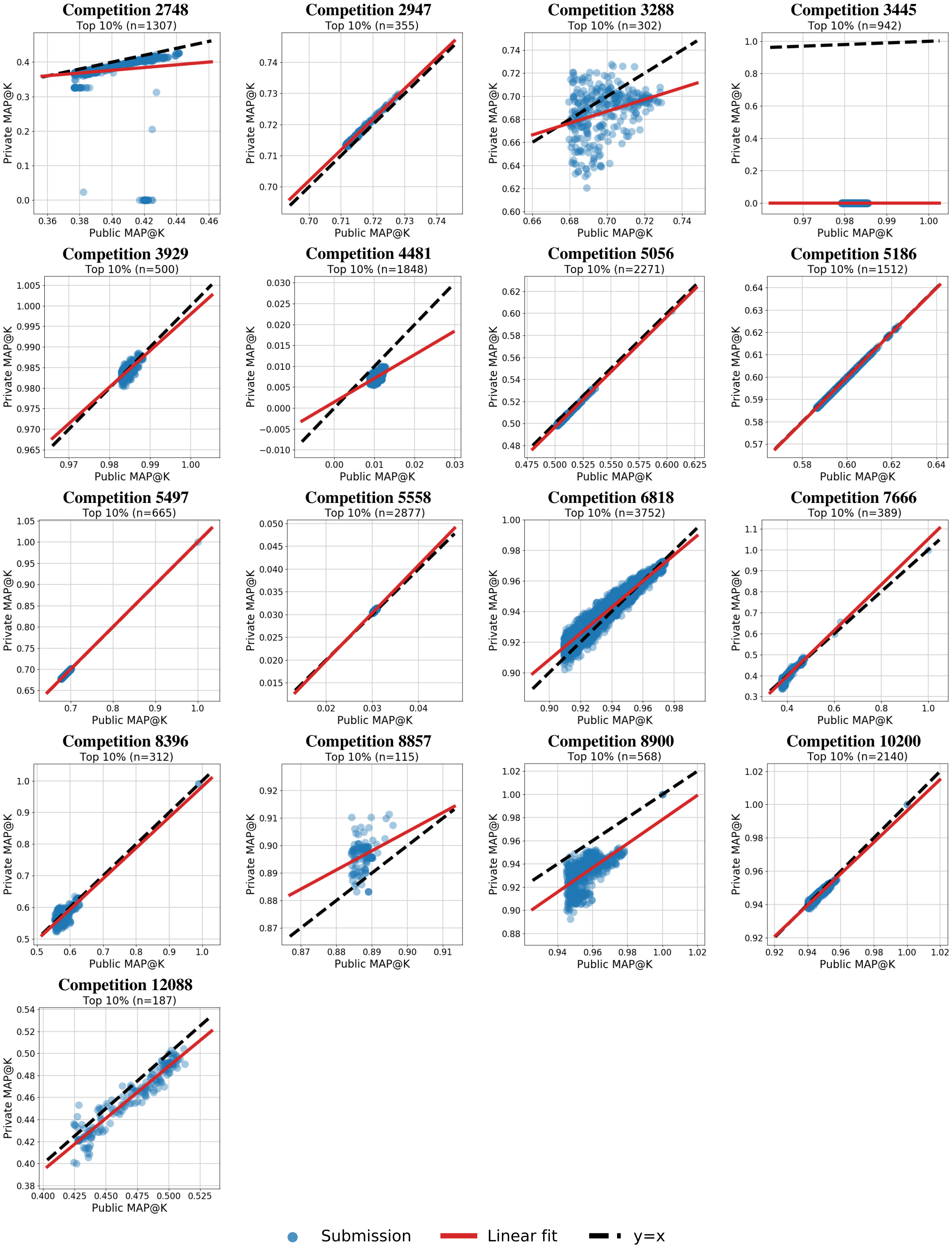

Competition 2748
Top 10% (n=1307)
Competition 2947
Top 10% (n=355)
Competition 3288
Top 10% (n=302)
Competition 3445
Top 10% (n=942)
Competition 3929
Top 10% (n=500)
Competition 4481
Top 10% (n=1848)
Competition 5056
Top 10% (n=2271)
Competition 5186
Top 10% (n=1512)
Competition 5497
Top 10% (n=665)
Competition 5558
Top 10% (n=2877)
Competition 6818
Top 10% (n=3752)
Competition 7666
Top 10% (n=389)
Competition 8396
Top 10% (n=312)
Competition 8857
Top 10% (n=115)
Competition 8900
Top 10% (n=568)
Competition 10200
Top 10% (n=2140)
Competition 12088
Top 10% (n=187)

Figure 16: Private versus public MAP@K for top 10% of submissions for Kaggle MAP@K competitions.

E MulticlassLoss

Our MulticlassLoss category includes competitions marked in the MetaKaggle dataset as MulticlassLoss and MulticlassLossOld, as these metrics are functionally equivalent.

E.1 MulticlassLoss: competition info

Table 5: Competitions scored with MulticlassLoss with greater than 1000 submissions. n_{public} is the size of the public test set and n_{private} is the size of the private test set.

MulticlassLoss				
ID	Name	# Sub.	n_{public}	n_{private}
3043	Predict Closed Questions on Stack Overflow	1,431	0	81,686
3978	National Data Science Bowl	15,121	9,128	121,273
4117	Microsoft Malware Classification Challenge (BIG 2015)	7,670	3,262	7,611
4280	Otto Group Product Classification Challenge	43,524	101,058	43,310
4521	Right Whale Recognition	4,789	2,424	4,501
4588	Telstra Network Disruptions	19,603	3,240	7,931
4654	Walmart Recruiting: Trip Type Classification	13,135	28,702	66,972
5048	State Farm Distracted Driver Detection	25,591	15,148	64,579
5340	TalkingData Mobile User Demographics	24,631	33,621	78,450
5568	The Nature Conservancy Fisheries Monitoring	2,100	1,132	13,021
5590	Two Sigma Connect: Rental Listing Inquiries	47,044	29,864	44,796
6243	Intel & MobileODT Cervical Cancer Screening	1,367	589	3,943
6841	Personalized Medicine: Redefining Cancer Treatment	3,056	2,267	3,401
7516	Spooky Author Identification	10,640	2,518	5,874

E.2 MulticlassLoss: outlier competitions

For MulticlassLoss, we investigated competitions whose linear fit was a bad approximation to the diagonal $y = x$ line. We first identified these competitions using visual inspection of the plots in Section E.3.

By reading the Kaggle forums and the description of the dataset construction, we were able to determine a possible cause of overfitting for most of the outlier competitions. In most cases, the competition hosts did not create the public and private test splits in a truly i.i.d. manner, and competitiors were able to exploit the difference in distribution. In other cases, the size of the test set was extremely small. The list of outlier competitions and corresponding cause of overfitting follow. We also include links to relevant descriptions of the data or discussions on the Kaggle forums.

- **Competition 3043: Predict Closed Questions on Stack Overflow** All test data was considered private, and the public leaderboard was identical to the private leaderboard. `https://www.kaggle.com/c/predict-closed-questions-on-stack-overflow/leaderboard`
- **Competition 5568: The Nature Conservancy Fisheries Monitoring** The test set was released in two stages. `https://www.kaggle.com/c/the-nature-conservancy-fisheries-monitoring/data`
- **Competition 6243: Intel & MobileODT Cervical Cancer Screening** The test set was released in two stages. `https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening/data`
- **Competition 6841: Personalized Medicine: Redefining Cancer Treatment** The test set was released in two stages because a team discovered the test data from Stage 1, including labels, on a public website. `https://www.kaggle.com/c/msk-redefining-cancer-treatment/discussion/36383#latest-223033`

E.3 MulticlassLoss: all submissions

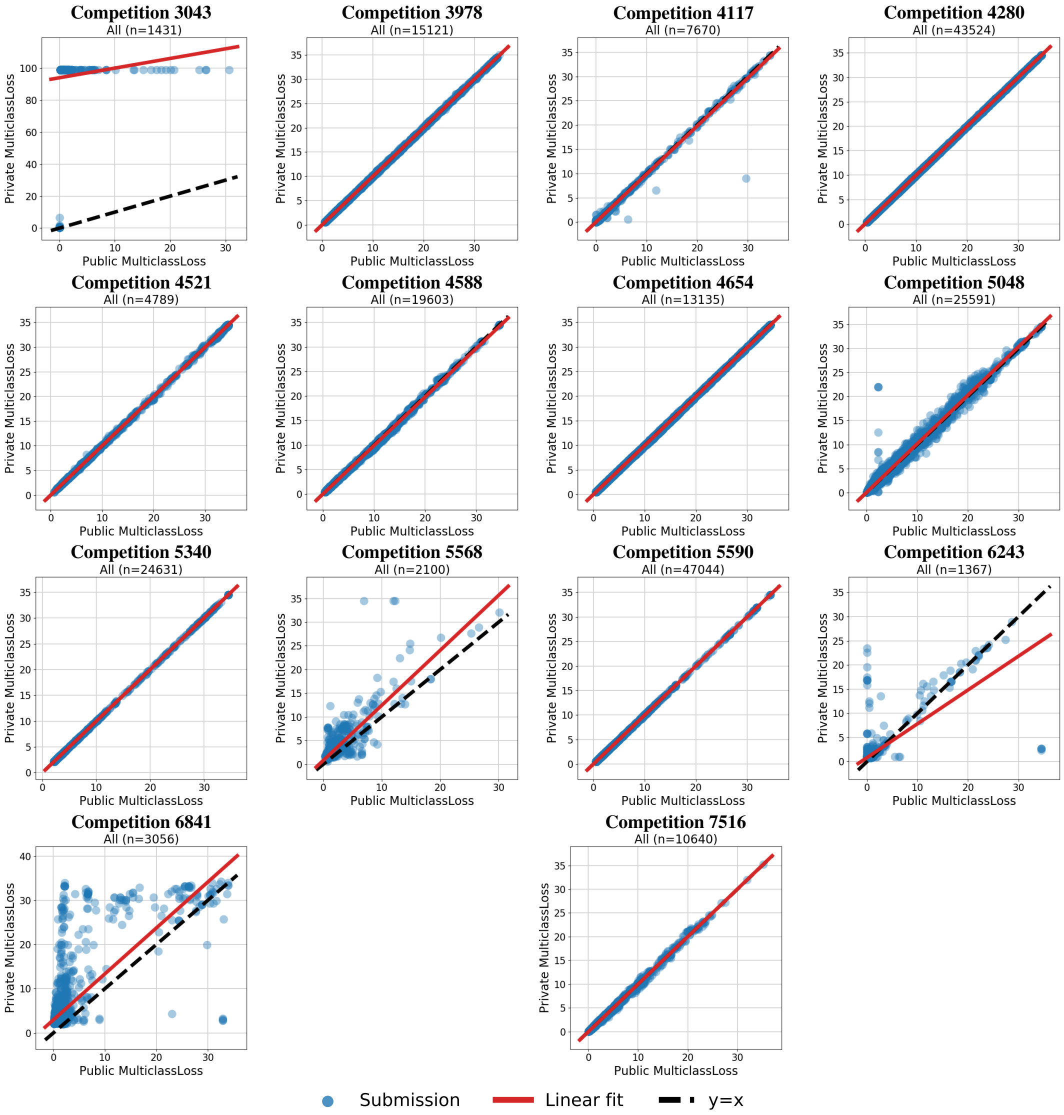


Figure 17: Private versus public MulticlassLoss for all submissions for Kaggle MulticlassLoss competitions.

E.4 MulticlassLoss: top 10% of submissions

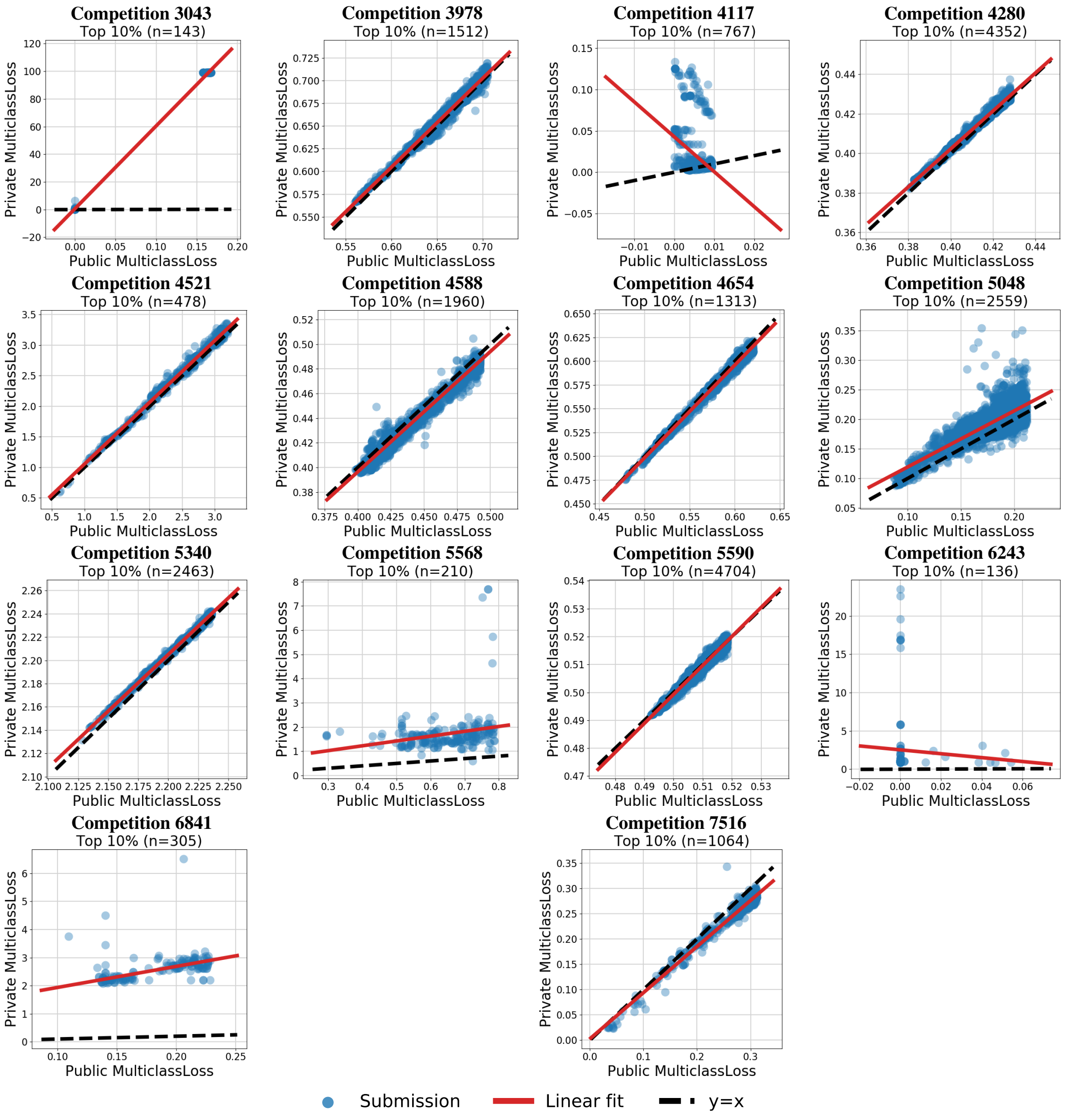


Figure 18: Private versus public MulticlassLoss for top 10% of submissions for Kaggle MulticlassLoss competitions.

F LogLoss

F.1 LogLoss: competition info

Table 6: Competitions scored with LogLoss with greater than 1000 submissions. n_{public} is the size of the public test set and n_{private} is the size of the private test set. N/A means that we could not access the competition data to compute the dataset sizes. A * after the competition name means the name was slightly edited to fit into the table.

LogLoss				
ID	Name	# Sub.	n_{public}	n_{private}
2780	Predicting a Biological Response	8,843	625	1,876
2984	Practice Fusion Diabetes Classification	2,200	1,245	3,734
3377	ICDAR2013 - Gender Prediction from Handwriting	1,884	286	486
3934	Display Advertising Challenge	8,640	1,208,427	4,833,708
3984	Tradeshift Text Classification	5,659	163,525	381,557
4120	Click-Through Rate Prediction	31,019	915,493	3,661,972
4438	Avito Context Ad Clicks	5,951	4,788,454	11,173,061
4852	BNP Paribas Cardif Claims Management	54,519	37,750	76,643
4862	March Machine Learning Mania 2016	1,054	68	2,210
6004	Data Science Bowl 2017	1,694	N/A	N/A
6277	Quora Question Pairs	53,927	140,748	2,205,048
7163	WSDM - KKBox's Churn Prediction Challenge	6,256	361,788	542,683
7380	Statoil/C-CORE Iceberg Classifier Challenge	42,936	674	7,750
7823	Attrition de clientes	1,911	15,000	15,000
8310	Google Cloud & NCAA® ML Competition 2018*	1,660	342	11,049

F.2 LogLoss: outlier competitions

For LogLoss, we observed one competition (Competition 6004: Data Science Bowl 2017) whose linear fit was a bad approximation the diagonal $y = x$ line. The competition shows approximately 13 outliers when looking at the scatter plot for the top 10% of submissions (see Figure 20).

The competition is unique in two aspects: first, the competition used two stages with separate test sets released at different times and second, participants were allowed to use external data. Participants were eventually given all the labels to the Stage 1 test data, but the Stage 2 test data was different in distribution from the Stage 1 test data (see `https://www.kaggle.com/c/data-science-bowl-2017/discussion/31383#latest-174302`). It is unclear whether the final private test data was generated from an i.i.d. split with either the Stage 1 test data or the Stage 2 data. Moreover, training the model on external data violates our assumption that the models were only trained and evaluated on i.i.d. data.

F.3 LogLoss: all submissions

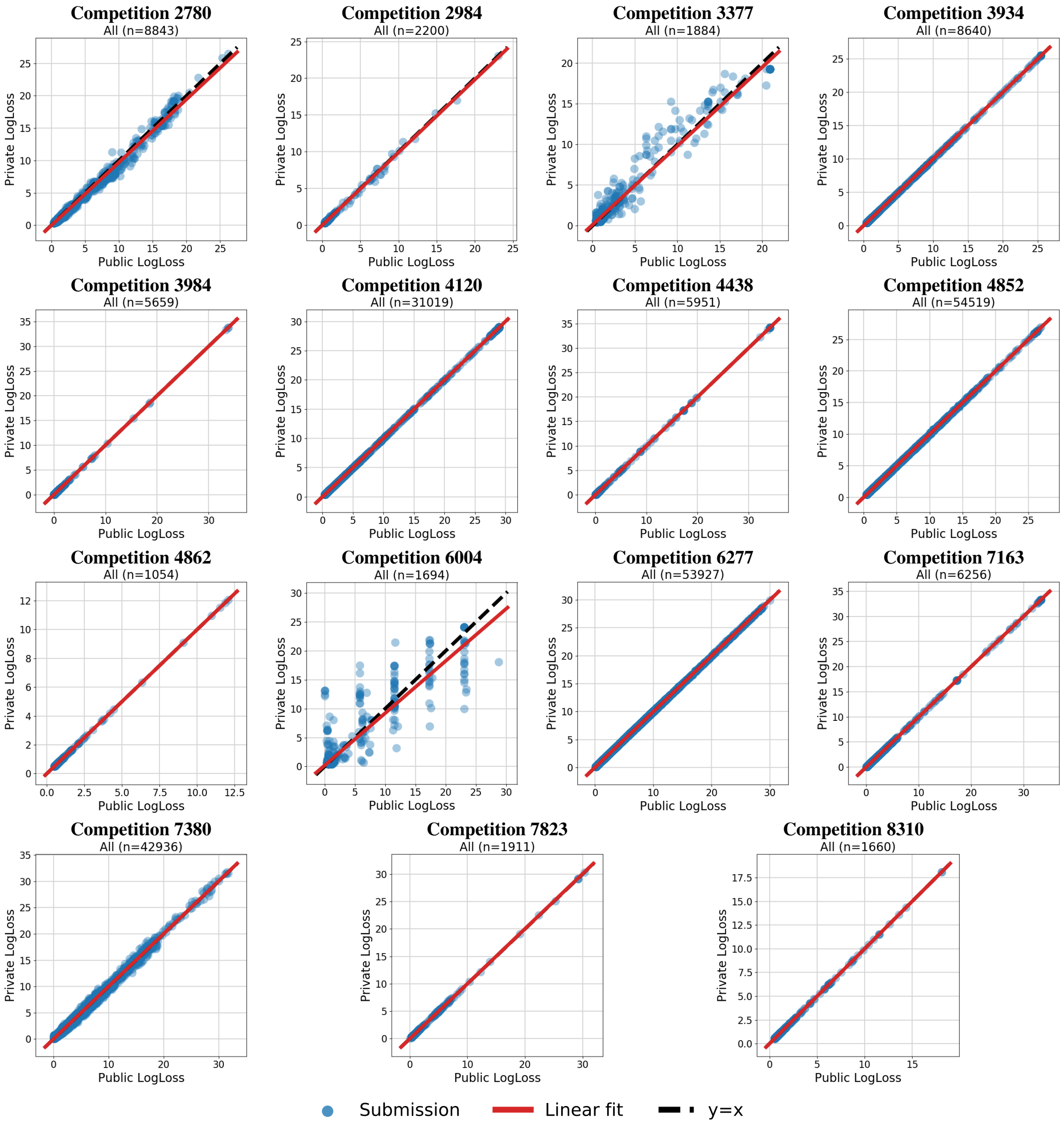


Figure 19: Private versus public LogLoss for all submissions for Kaggle LogLoss competitions.

F.4 LogLoss: top 10% of submissions

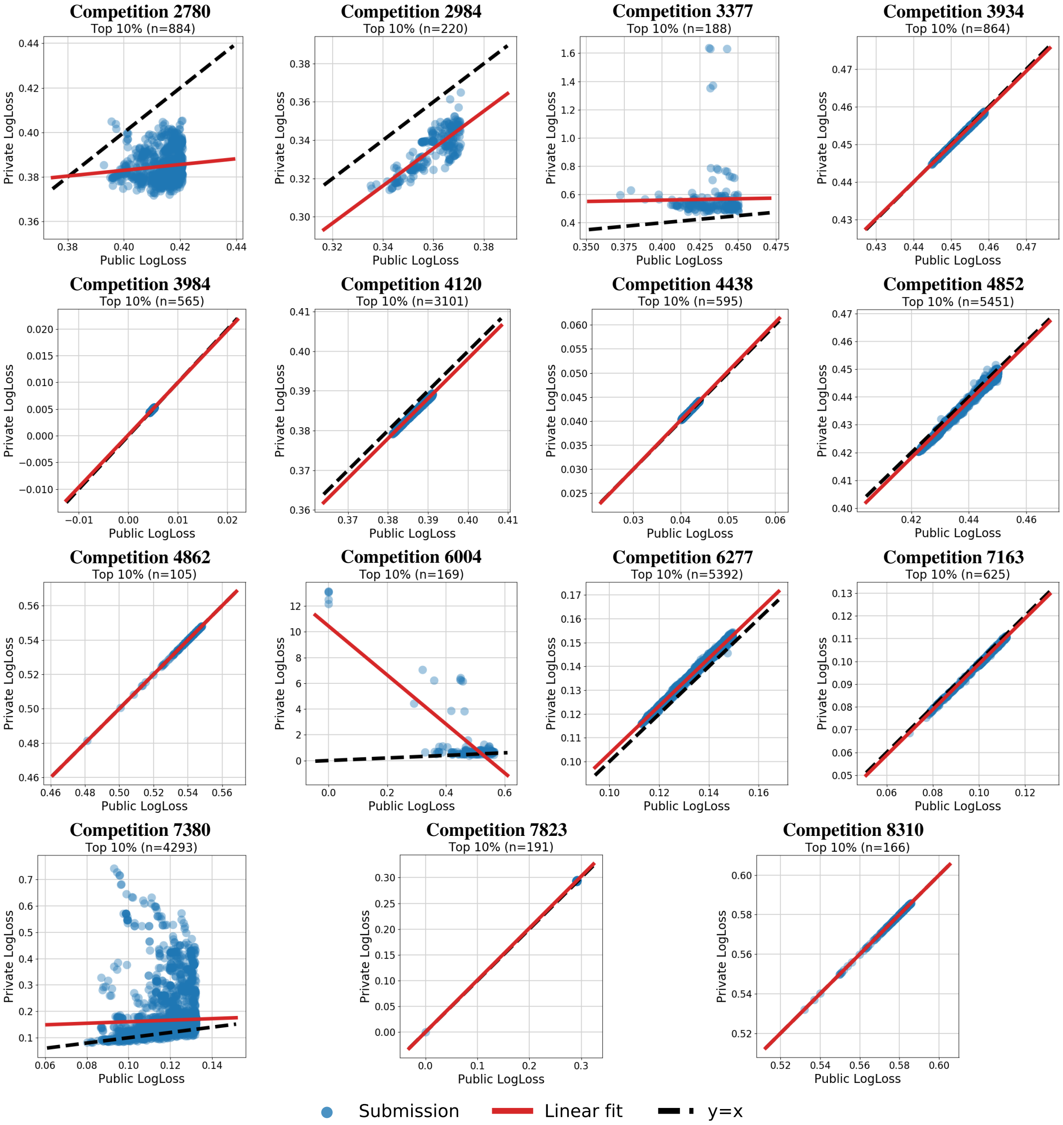


Figure 20: Private versus public LogLoss for top 10% of submissions for Kaggle LogLoss competitions.

G Mean score differences over time

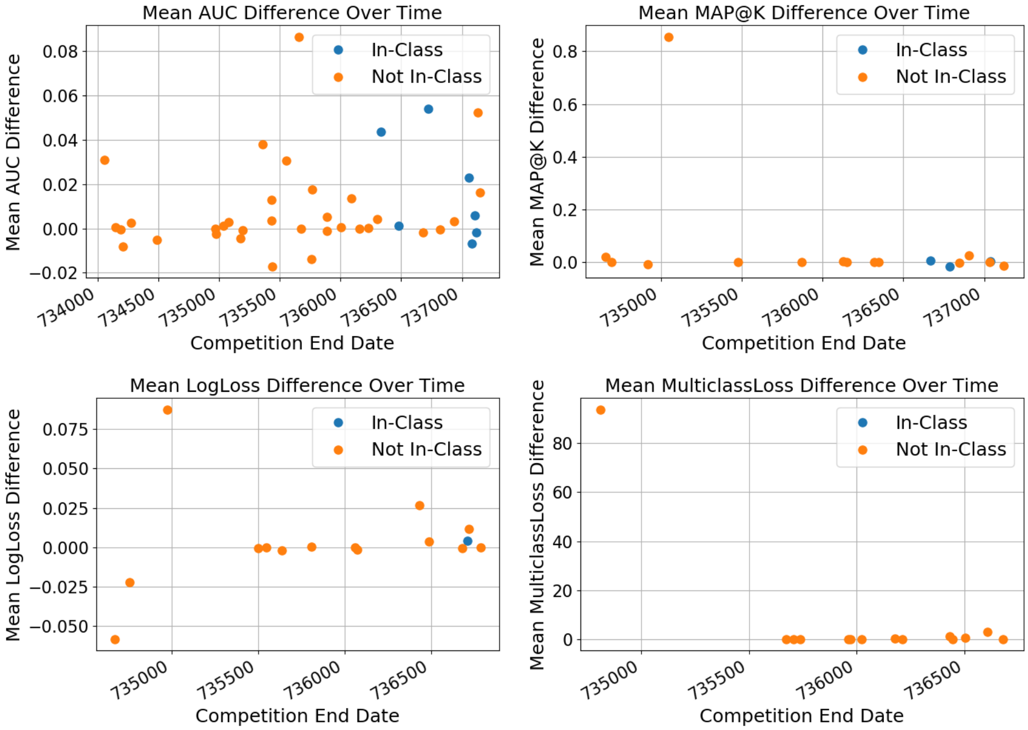


Figure 21: Mean score differences (across all pre-deadline submissions) versus competition end date for all classification evaluation metrics with at least 10 competitions with at least 1000 submissions each. Score difference is public minus private for metrics where a higher score is better (AUC, CategorizationAccuracy (included in the main text), and MAP@K) and private minus public for metrics where a lower score is better (LogLoss and MulticlassLoss). Blue dots show InClass competitions and orange dots show all other competition types (Featured, Research, Getting Started, and Playground).