Competition over data: when does data purchase benefit users?

Anonymous Authors[1]

Abstract

As machine learning (ML) is deployed by many competing service providers, the underlying ML predictors also compete against each other, and it is increasingly important to understand the impacts and biases from such competition. In this paper, we study what happens when the competing predictors can acquire additional labeled data to improve their prediction quality. We introduce a new environment that allows ML predictors to use active learning algorithms to purchase labeled data within their budgets while competing against each other to attract users. Our environment models a critical aspect of data acquisition in competing systems which has not been well-studied before. When predictors can purchase additional labeled data, their overall performance improves. Surprisingly, however, the quality that users experience—i.e. the accuracy of the predictor selected by each user—can decrease even as the individual predictors get better. We show that this phenomenon naturally arises due to a trade-off whereby competition pushes each predictor to specialize in a subset of the population while data purchase has the effect of making predictors more uniform.

1. Introduction

It is becoming increasingly common for companies to make use of machine learning (ML) predictions in their services (Linden et al., 2003; Covington et al., 2016; Marr, 2019; Nanduri et al., 2020). When there are many comparable ML-based services in the same market, customers usually compare multiple services and choose only one company that provides the best service. This user selection creates competition between companies. In the competition, to attract customers, the competing companies strive to produce high-quality ML predictions, often buying customer data or subscriptions to data marketplaces (Meierhofer et al., 2019). For example, the companies State Farm, Progressive, and AllState are competitors that use ML predictions to analyze user behavior, assess risk, and adjust premiums in the U.S. auto insurance market (Sennaar, 2019). They offer the Pay-How-You-Drive type of insurance, and customers with this insurance pay lower premiums than regular auto insurance on the condition that the insurer monitors driving patterns such as rapid acceleration or oscillations in speed (Arumugam and Bhargavi, 2019; Jin and Vasserman, 2019). In other words, auto insurance companies provide financial benefits to customers for the purpose of collecting driving pattern data. They use the purchased data for their business (e.g. updating the insurance recommendation model or reassessing the risks) while competing with each other.

Analyzing the effects of data purchase in competitions could have practical implications, but it has not been considered much in the ML literature. The effects of data acquisition have been studied extensively in the field of active learning (AL), which is the problem of finding effective data to label (Settles, 2009; Ren et al., 2020). However, since the AL usually assumes a single agent situation, it is not straightforward to establish competing systems that require more than one competitor. Recently, Ginart et al. (2021) studied the impacts of competitions by modeling an environment where several predictors compete for user data. They showed that competition pushes competing predictors to focus on a subset of the population and helps users find high-quality predictions. It describes interesting implications of competition, but their model did not allow predictors to purchase data, which is the focus of our work. Related works are further discussed in Appendix.

Contributions In this paper, we study what happens when competing predictors purchase customer data to improve their ML models. As a summary, our main contributions are as follows.

- We propose a novel environment that can models various real-world competitions. Our environment allows ML predictors to use AL algorithms to purchase labeled data within a finite budget while competing against each other (Sec. 2).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

- Surprisingly, our results show that when competing predictors purchase data, the quality of the predictions selected by each user can decrease even as competing ML predictors get better (Sec. 3.1).

- We explain this counterintuitive finding by demonstrating that data purchase makes competing predictors similar to each other. Thus the users' available options are reduced (Sec. 3.2).

2. A general environment for competition and data purchase

This section formally introduces a new and general competitive environment that allows competing predictors to acquire data points within a finite budget. In our environment, competition is represented by a series of interactions between a sequence of users and the fixed number of competing predictors, where the interaction is modeled by supervised learning tasks. To be more specific, we define some notations.

Notations At each round $t \in [T] := \{1, \dots, T\}$, we denote a user query by $X_t \in \mathcal{X}$ and its associated user label by $Y_t \in \mathcal{Y}$. Throughout this paper, we focus on classification problems, *i.e.*, $|\mathcal{Y}|$ is finite, while our environment can easily extend to regression cases. We denote a user stream by $\{(X_t, Y_t)\}_{t=1}^T$ and assume users are independent and identically distributed (i.i.d.) by some distribution $P_{X,Y}$. We call $P_{X,Y}$ the user distribution. As for the predictor side, we suppose there are M competing predictors. For $i \in [M]$, each ML predictor is defined as a tuple $\mathcal{C}^{(i)} := (n_{\mathrm{s}}^{(i)}, n_{\mathrm{b}}^{(i)}, f^{(i)}, \pi^{(i)})$, where $n_{\mathrm{s}}^{(i)} \in \mathbb{N}$ is the number of i.i.d. seed data points from $P_{X,Y}$, $n_{\mathrm{b}}^{(i)} \in \mathbb{N}$ is a budget, $f^{(i)} : \mathcal{X} \to \mathcal{Y}$ is an ML model, and $\pi^{(i)} : \mathcal{X} \to \{0, 1\}$ is a buying strategy. We consider a predictor $\mathcal{C}^{(i)}$ initially owns $n_{\mathrm{s}}^{(i)}$ data points and can additionally purchase user data within $n_{\mathrm{b}}^{(i)}$ budgets. With abuse of notation, we allow the model $f^{(i)}$ and buying strategy $\pi^{(i)}$ to be updated throughout the T competition rounds.

Competition dynamics Before the first round, all the M competing predictors independently train their model $f^{(i)}$ using the $n_{\mathrm{s}}^{(i)}$ seed data points. After this, at each round $t \in [T]$, a user sends a query X_t to all the predictors $\{\mathcal{C}^{(j)}\}_{j=1}^M$, and each predictor $\mathcal{C}^{(i)}$ determines whether to buy the user data. We describe this decision by using the buying strategy $\pi^{(i)}$. If the predictor $\mathcal{C}^{(i)}$ thinks that the labeled data would be worth one unit of budgets, we denote this by $\pi^{(i)}(X_t) = 1$. Otherwise, if $\mathcal{C}^{(i)}$ thinks that it is not worth one unit of budgets, then $\pi^{(i)}(X_t) = 0$. As for the $\pi^{(i)}$, we allow ML predictors to use any stream-based AL algorithm (Freund et al., 1997; Žliobaitė et al., 2013). For example, a predictor $\mathcal{C}^{(i)}$ can use the uncertainty-based AL rule (Settles and Craven, 2008); $\mathcal{C}^{(i)}$ considers purchasing user data if the current prediction $f^{(i)}(X_t)$ is not certain (e.g. the Shannon's entropy of $p^{(i)}(X_t)$ is higher than some predefined threshold value where $p^{(i)}(X_t)$ is the corresponding probability estimate for $f^{(i)}(X_t)$). In brief, we suppose a predictor $\mathcal{C}^{(i)}$ shows purchase intent if the remaining budget is greater than zero and $\pi^{(i)}(X_t) = 1$. If the remaining budget is zero or $\pi^{(i)}(X_t) = 0$, then $\mathcal{C}^{(i)}$ provides a prediction $f^{(i)}(X_t)$ to the user.

We now elaborate on how a user selects one predictor. We assume that the user selects only one predictor based on both purchase intents and prediction information received from $\{\mathcal{C}^{(j)}\}_{j=1}^M$. If there are more than or equal to one buyer, then a user selects one of the buyers uniformly at random. We can think that users prioritize financial advantages (e.g. discounts) over the prediction quality and do not care much about a particular company when there are many buyers. Once selected, only the selected predictor's budget is reduced by one; all other predictor's budget stays the same. If no predictor shows purchase intent and receives prediction information $\{f^{(j)}(X_t)\}_{j=1}^M$, then a user chooses the predictor $\mathcal{C}^{(i)}$ with the following probability.

$$P\left(W_t = i \mid Y_t, \{f^{(j)}(X_t)\}_{j=1}^M\right) = \frac{\exp\left(\alpha q\left(Y_t, f^{(i)}(X_t)\right)\right)}{\sum_{j=1}^M \exp\left(\alpha q\left(Y_t, f^{(j)}(X_t)\right)\right)}, \tag{1}$$

where α denotes a temperature parameter and $q : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is a predefined quality function that measures similarity between the user label Y_t and the prediction. We assume that users are *rational* in that users are more likely to select high-quality predictions, *i.e.,* $\alpha \geq 0$.

Afterwards, we denote the index of selected predictor by $W_t \in [M]$. The selected predictor $\mathcal{C}^{(W_t)}$ gets the user label Y_t and updates the model $f^{(W_t)}$ by training on (X_t, Y_t). The other predictors $f^{(i)}$ stay the same for $i \neq W_t$. We describe our competition system in Appendix.

Example 1 (Auto insurance in Sec. 1)**.** *X_t is the t-th driver's demographic information and driving history, and Y_t is the driver's preferred insurance plan. Each predictor $\mathcal{C}^{(i)}$ is one insurance company (e.g., State Farm, Progressive, or AllState). In regular competition, each company $\mathcal{C}^{(i)}$ offers an auto insurance plan $f^{(i)}(X_t)$ based on what it predicts to be most suitable for this driver. The driver chooses one company whose offered plan is the closest to Y_t. If a company believes that in its database there are infrequent data from a particular group of drivers t-th driver belongs to (e.g. new drivers in their 30s), it can offer this driver discounts to attract her and collect her driving pattern data. The acquired data is then used to improve the company's future predictions.*

Characteristics of our environment Our environment simplifies data purchase and real-world competition, which usually exist in much more complicated forms, yet it has several powerful characteristics. Our environment is realistic in that it represents the rational preference of customers and companies in data purchase. Customers want to choose the best service after comparing options but can choose another service when there are financial benefits. As for the companies, which always try to provide the best services, they can attract users with promotional coupons, discounts, or free services (Rowley, 1998; Reimers and Shiller, 2019). Although this could be costly for the company, it enables them to collect user data points, make use of the purchased data, and improve their ML predictors.

3. Experiments

Using the proposed environment, we investigate the impacts of the data purchase on the quality and diversity of predictions across different user distributions. Our experiments show an interesting phenomenon that data purchase can decrease the quality of the predictor selected by a user, even when the quality of the predictors gets improved on average (Sec. 3.1). In addition, we demonstrate that data purchase makes ML predictors similar to each other. Data purchase reduces the effective variety of options (Sec. 3.2).

Metrics To quantitatively measure the effects of data purchase, we introduce evaluation metrics. First, we define the overall quality as $\mathbb{E}\left[\frac{1}{M}\sum_{j=1}^{M} q\left(Y, f^{(j)}(X)\right)\right]$, where the expectation is taken over the user distribution $P_{X,Y}$. The overall quality represents the average quality that competing predictors provide in the market. Another type of quality metric is the quality of experience (QoE), the quality of the predictor selected by a user. The QoE is defined as $\mathbb{E}\left[q\left(Y, f^{(W)}(X)\right)\right]$. Here, $W \in [M]$ is a random variable for a selected index defined in Equation (1), and the expectation is considered over the random variables (X, Y, W). Considering that a user selects one predictor based on Equation (1), QoE can be considered as the utility of users.

Next, we define the diversity to quantify how variable the ML predictions are. To be more specific, for $i \in \mathcal{Y}$, let $p_i = \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}(f^{(j)}(X) = i)$ be the proportion of predictors whose prediction is i. The diversity is defined as $\mathbb{E}\left[-\sum_{i\in\mathcal{Y}} p_i \log(p_i)\right]$, where the expectation is taken over the marginal distribution P_X and we use the convention $0\log(0) = 0$ when $p_i = 0$. Note that the diversity is defined as the expected Shannon's entropy of competing ML predictions. When there are various different options that a user can choose from, the diversity is more likely to be large.

In our experiments, we evaluate all the metrics after the T competition rounds finish, and thus we do not perform data buying procedures during the evaluation. As for the evaluation, since it is difficult to compute exact expectations, we estimate them with the sample averages using the held-out test data that are not used during the competition rounds.

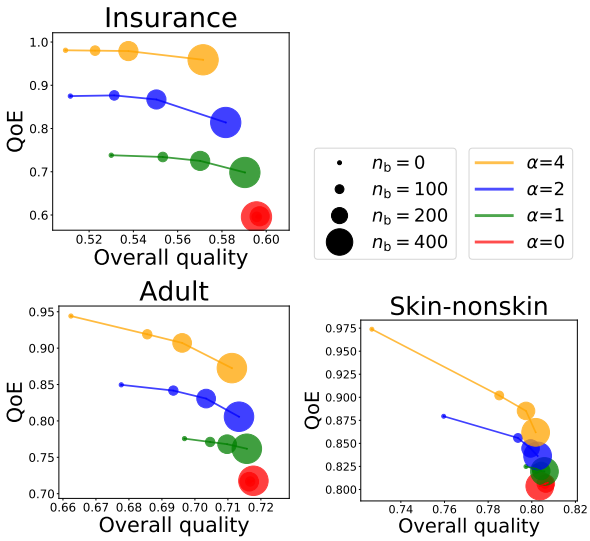

Figure 1. The illustrations of QoE as a function of the overall quality. In several settings, the overall quality increases as more budgets are used, but QoE decreases.

Implementation protocol Our experiments consider the seven real datasets for $P_{X,Y}$, namely `Insurance` (Van Der Putten and van Someren, 2000), `Adult` (Dua and Graff, 2017), and `Skin-nonskin` (Chang and Lin, 2011) datasets. Throughout the experiments, we fix the total number of competition rounds to $T = 10^4$, the number of predictors to $M = 18$, and a quality function to the correctness function, *i.e.*, $q(Y_1, Y_2) = \mathbb{1}(\{Y_1 = Y_2\})$ for all $Y_1, Y_2 \in \mathcal{Y}$. We consider various competition situations by varying the budget $n_b \in \{0, 100, 200, 400\}$ and the temperature parameter $\alpha \in \{0, 1, 2, 4\}$. For each pair (n_b, α), which generates one competition environment, we repeat experiments 30 times to obtain stable estimates for the metrics. We provide more details on implementations in Appendix.

3.1. Effects of data purchase on quality

We first study how data purchase affects the overall quality and the QoE in various competition settings. Fig. 1 illustrates how QoE changes with respect to (w.r.t.) the overall quality. When $\alpha > 0$, data purchase increases the overall quality as n_b increases across all datasets. This can be explained as follows. Given that a predictor buys user data using a stream-based AL algorithm (e.g., a predictor buys when its prediction is highly uncertain), the active data ac-

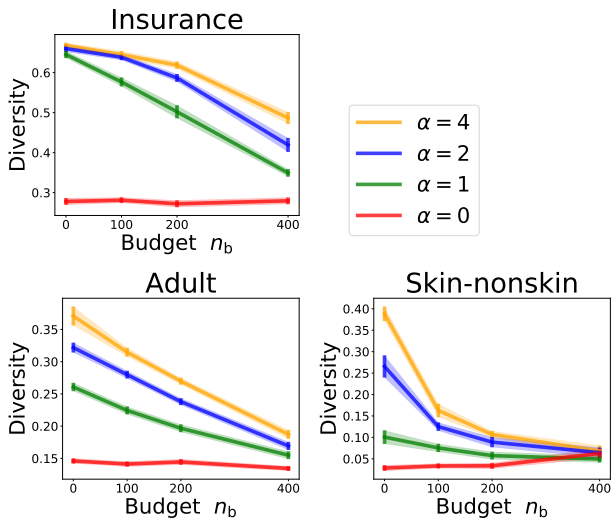

Figure 2. The illustration of the diversity as a function of the budget n_b. We denote a 99% confidence band based on 30 independent runs. Competing ML predictors become similar in a sense that the diversity decreases as the budget increases.

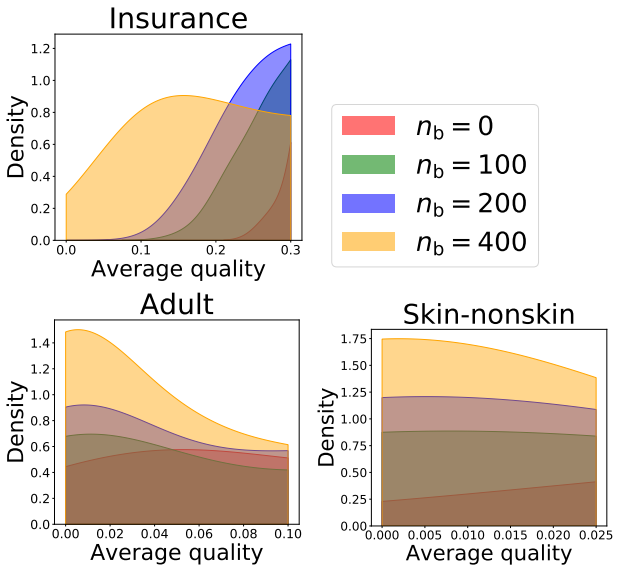

Figure 3. Probability density plots of the average quality $\frac{1}{M}\sum_{j=1}^{M} q(Y, f^{(j)}(X))$ at near zero. The probability that all ML predictors produce low-quality prediction at the same time increases, and users might not be satisfied with the ML predictions after the competing predictors purchase data.

quisition reduces the model's uncertainty and increases the quality of the individual model. As a result, data purchase increases the overall quality. As for the QoE, however, data purchase mostly decreases QoE as n_b increases. For example, when the user distribution is `Insurance` and $\alpha = 2$, QoE is 0.875 when $n_b = 0$, but it reduces to 0.814 when $n_b = 400$, which correspond to 7% reduction.

Implications of data purchase on quality As Fig. 1 shows, in most cases, surprisingly, QoE decreases even when the overall quality increases. In other words, the quality that competing predictors provide is generally improved, but it does not necessarily mean that users will be more satisfied with the ML predictions. Although this result might sound counterintuitive, we argue that it could happen if the data purchase makes users experience fewer options and increases the probability of finding low-quality predictions when n_b increases. To verify this, in the next section, we investigate how data purchase affects the diversity.

3.2. Effects of data purchase on diversity

We now study how data purchase affects the diversity. Fig. 2 illustrates the diversity as a function of the budget n_b in various competition settings. When $\alpha > 0$, the diversity monotonically decreases w.r.t. n_b across all datasets. That is, the competing predictors get similar as more budgets are allowed. In particular, when $\alpha = 4$ and the dataset is `Adult`, the diversity is 0.371 on average when $n_b = 0$, but it reduces to 0.187, which correspond to 50% reduction, when $n_b = 400$.

Implications of data purchase on diversity As shown in Sec. 3.1, data purchase can hurt the quality of the predictor selected by a user. To explain this phenomenon, we demonstrate that the probability of finding low-quality prediction increases due to the reduction in the diversity. We illustrate the probability density functions of the average quality near zero in Fig. 3. It clearly shows that the probability that the average quality is near zero increases as more budgets n_b are used: The areas for $n_b = 400$ (colored in yellow) are clearly larger than those for $n_b = 0$ (colored in red). As predictions become similar, it is more likely that all ML predictions are poor at the same time, and thus the probability that users are not satisfied with the predictions increases.

4. Conclusion

In this paper, characterizing the nature of competition and data purchase, we propose a new competitive environment that allows predictors to actively acquire user labels. Our results show that even though the data purchase improves the quality that predictors provide, it can decrease the quality that users experience. We explain this counterintuitive finding by demonstrating that data purchase makes competing predictors similar to each other.

References

Arumugam, S. and Bhargavi, R. (2019). A survey on driving behavior analysis in usage based insurance using big data.

Journal of Big Data, 6(1):1–21.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Covington, P., Adams, J., and Sargin, E. (2016). Deep neural networks for youtube recommendations. In *Proceedings of the 10th ACM conference on recommender systems*, pages 191–198.

Dua, D. and Graff, C. (2017). Uci machine learning repository.

Freund, Y., Seung, H. S., Shamir, E., and Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine learning*, 28(2):133–168.

Ginart, T., Zhang, E., Kwon, Y., and Zou, J. (2021). Competing ai: How does competition feedback affect machine learning? In *International Conference on Artificial Intelligence and Statistics*, pages 1693–1701. PMLR.

Jin, Y. and Vasserman, S. (2019). Buying data from consumers: The impact of monitoring programs in us auto insurance. *Unpublished manuscript. Harvard University, Department of Economics, Cambridge, MA*.

Linden, G., Smith, B., and York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80.

Marr, B. (2019). The amazing ways ebay is using artificial intelligence to boost business success. `https://www.forbes.com/sites/bernardmarr/2019/04/26/the-amazing-ways-ebay-is-using-artificial-intelligence-to-boost-business-success`. Posted April 26, 2019; Retrieved May 19, 2021.

Meierhofer, J., Stadelmann, T., and Cieliebak, M. (2019). Data products. In *Applied Data Science*, pages 47–61. Springer.

Nanduri, J., Jia, Y., Oka, A., Beaver, J., and Liu, Y.-W. (2020). Microsoft uses machine learning and optimization to reduce e-commerce fraud. *INFORMS Journal on Applied Analytics*, 50(1):64–79.

Reimers, I. and Shiller, B. R. (2019). The impacts of telematics on competition and consumer behavior in insurance. *The Journal of Law and Economics*, 62(4):613–632.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Chen, X., and Wang, X. (2020). A survey of deep active learning. *arXiv preprint arXiv:2009.00236*.

Rowley, J. (1998). Promotion and marketing communications in the information marketplace. *Library review*.

Sennaar, K. (2019). How america's top 4 insurance companies are using machine learning. `https://emerj.com/ai-sector-overviews/machine-learning-at-insurance-companies`. Posted February 26, 2020; Retrieved May 19, 2021.

Settles, B. (2009). Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

Settles, B. and Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079.

Van Der Putten, P. and van Someren, M. (2000). Coil challenge 2000: The insurance company case. Technical report, Technical Report 2000–09, Leiden Institute of Advanced Computer Science.

Žliobaitė, I., Bifet, A., Pfahringer, B., and Holmes, G. (2013). Active learning with drifting streaming data. *IEEE transactions on neural networks and learning systems*, 25(1):27–39.