A. Proofs

For the proofs, we make use of the following result from Jiang (2019) which bounds the number of distinct k-NN sets on the sample across all k:

Lemma 1 (Lemma 3 of Jiang (2019))**.** *Let M be the number of distinct k-NN sets over $\mathcal{X}$, that is, $M := |\{N_k(x) : x \in \mathcal{X}\}|$. Then $M \leq D \cdot n^D$.*

Proof of Theorem 1. We have by triangle inequality and the smoothness condition in Assumption 1 that:

$$\begin{aligned}|\eta_k(x) - \eta(x)| &\leq \left|\sum_{i=1}^n (\eta(x_i) - \eta(x)) \cdot \frac{1\left[x_i \in N_k(x)\right]}{|N_k(x)|}\right| + \left|\sum_{i=1}^n (y_i - \eta(x_i)) \cdot \frac{1\left[x_i \in N_k(x)\right]}{|N_k(x)|}\right| \\ &\leq C_\alpha \cdot r_k(x)^\alpha + \left|\sum_{i=1}^n (y_i - \eta(x_i)) \cdot \frac{1\left[x_i \in N_k(x)\right]}{|N_k(x)|}\right|.\end{aligned}$$

We now bound each of the two terms separately.

To bound $r_k(x)$, let $r = \left(\frac{2k}{\omega \cdot v_D \cdot n \cdot p_{X,0}}\right)^{1/D}$. We have $\mathcal{P}(B(x,r)) \geq \omega \inf_{x' \in B(x,r) \cap \mathcal{X}} p_X(x') \cdot v_D r^D \geq \omega p_{X,0} v_D r^D = \frac{2k}{n}$, where $\mathcal{P}$ is the distribution function w.r.t. p_X. By Lemma 7 of Chaudhuri & Dasgupta (2010) and the condition on k, it follows that with probability $1 - \delta/2$, uniformly in $x \in \mathcal{X}$, $|B(x,r) \cap X| \geq k$, where X is the sample of feature vectors. Hence, $r_k(x) < r$ for all $x \in \mathcal{X}$ uniformly with probability at least $1 - \delta/2$.

Define $\xi_i := y_i - \eta(x_i)$. Then, we have that $-1 \leq \xi_i \leq 1$ and thus by Hoeffding's inequality, we have that $A_x := \sum_{i=1}^n (y_i - \eta(x_i)) \cdot \frac{1[x_i \in N_k(x)]}{|N_k(x)|} = \sum_{i=1}^n \xi_i \cdot \frac{1[x_i \in N_k(x)]}{|N_k(x)|}$ satisfies $P(|A_x| > t/k) \leq 2\exp\left(-t^2/2k\right)$. Then setting $t = \sqrt{2k} \cdot \sqrt{\log(4D/\delta) + D\log(n)}$ gives

$$\mathbb{P}\left(|A_x| \geq \sqrt{\frac{2\log(4D/\delta) + 2D\log(n)}{k}}\right) \leq \frac{\delta}{2D \cdot n^D}.$$

By Lemma 3 of Jiang (2019), the number of unique random variables A_x across all $x \in \mathcal{X}$ is bounded by $D \cdot n^D$. Thus, by union bound,

$$\mathbb{P}\left(\sup_{x \in X} |A_x| \geq \sqrt{\frac{2\log(4D/\delta) + 2D\log(n)}{k}}\right) \leq \delta/2.$$

The result follows. □

Proof of Theorem 2. Let X be the n sampled feature vectors and let $x \in \mathcal{X}$. Define $k'(x) := |X \cap B(x, r_\beta(x))|$. We have:

$$|\eta_k(x) - \widetilde{\eta}_\beta(x)| \leq |\eta_{k'(x)}(x) - \eta_k(x)| + |\eta_{k'(x)}(x) - \widetilde{\eta}_\beta(x)|.$$

We bound each of the two terms separately. We have

$$|k'(x) - k| = \left|\sum_{x \in X} 1[x \in B(x, r(x))] - \beta \cdot n\right|$$

By Hoeffding's inequality we have

$$\mathbb{P}(|k'(x) - k| \geq t \cdot n) \leq 2\exp(-2t^2 n).$$

Choosing $t = \sqrt{\frac{\log(4D/\delta) + D\log(n)}{2n}}$ gives us

$$\mathbb{P}\left(|k'(x) - k| \geq \sqrt{\frac{n}{2} \cdot (\log(4D/\delta) + D\log(n))}\right) \leq \frac{\delta}{2D \cdot n^D}.$$

Dataset (m=5)	Accuracy (%)	Churn (%)	Churn Correct	Churn Incorrect
SVHN	90.34 (0.31)	6.61 (0.19)	2.75 (0.28)	43.12 (1.49)
MNIST	98.5 (0.07)	0.94 (0.14)	0.44 (0.09)	33.74 (4.39)
Fashion MNIST	89.71 (0.12)	4.05 (0.14)	1.85 (0.05)	23.16 (1.29)
CelebA Smiling	90.56 (0.09)	3.35 (0.16)	1.82 (0.11)	17.95 (0.99)
CelebA High Cheekbone	85.12 (0.16)	4.95 (0.2)	2.87 (0.1)	16.81 (1.24)
Phishing	96.11 (0.06)	0.54 (0.08)	0.29 (0.08)	6.77 (1.31)

Table 2. Ensemble results for all datasets. In all settings, the optimal m (number of subnetworks) is 5. We see that compared to the other methods presented, ensembling does well in both predictive performance and in reducing churn. It does come at a cost, however: the model is effectively 5 times larger, making both training and inference more expensive.

By Lemma 3 of Jiang (2019), the number of unique sets of points consisting of balls intersected with the sample is bounded by $D \cdot n^D$ and thus by union bound, we have with probability at least $1 - \delta/2$:

$$\sup_{x \in \mathcal{X}} |k'(x) - k| \leq \sqrt{\frac{n}{2} \cdot (\log(4D/\delta) + D\log(n)}).$$

We now have

$$\begin{aligned} |\eta_{k'(x)}(x) - \eta_k(x)| &\leq \left| \frac{1}{k} - \frac{1}{k'(x)} \right| \min\{k, k'(x)\} + \min\left\{\frac{1}{k}, \frac{1}{k'(x)}\right\} |k - k'(x)| \\ &\leq \frac{2}{k} \cdot |k - k'(x)| \leq \sqrt{\frac{2\log(4D/\delta) + 2D\log(n)}{\beta \cdot n}}. \end{aligned}$$

where the first inequality follows by comparing the difference contributed by the shared neighbors among the k-NN and $k'(x)$-NN (first term on RHS) and contributed by the neighbors that are not shared (second term on RHS).

For the second term, define $A_x := X \cap B(x, r_\beta(x))$. For any x' sampled from $B(x, r_\beta(x))$, we have that the expected label is $\widetilde{\eta}_\beta(x)$. Since $\eta_{k'(x)}(x)$ is the mean label among datapoints in A_x, then we have by Hoeffding's inequality that

$$\mathbb{P}(|\eta_{k'}(x) - \widetilde{\eta}_\beta(x)| \geq k'(x) \cdot t) \leq 2\exp\left(-t^2/2k'\right).$$

Then setting $t = \sqrt{2k'} \cdot \sqrt{\log(4D/\delta) + D\log(n)}$ gives

$$\mathbb{P}\left(|\eta_{k'(x)}(x) - \widetilde{\eta}_\beta(x)| \geq \sqrt{\frac{2\log(4D/\delta) + 2D\log(n)}{k'(x)}}\right) \leq \frac{\delta}{2D \cdot n^D}.$$

By Lemma 3 of Jiang (2019), the number of unique sets A_x across all $x \in \mathcal{X}$ is bounded by $D \cdot n^D$. Thus, by union bound, with probability at least $1 - \delta/2$L

$$|\eta_{k'(x)}(x) - \widetilde{\eta}_\beta(x)| \leq \sqrt{\frac{2\log(4D/\delta) + 2D\log(n)}{k'(x)}}.$$

The result follows immediately for n sufficiently large. □

B. Ensemble Results

In Table 2 we present the experimental results for the ensemble baseline. The method performs remarkably well, beating the proposed method and the other baselines on both accuracy and churn reduction across datasets. We do note, however, that ensembling does come at a cost which may prove prohibitive in many practical applications. Firstly, having m times the number of trainable parameters, training time (if done sequentially) takes m times as long, as does inference, since each subnetwork must be evaluated before aggregation.

Fixed	Ablated	Accuracy (%)	Churn (%)	Churn Correct
k = 10, a = 1	b = 0	86.54 (0.67)	13.43 (0.58)	5.86 (0.57)
	b = 0.05	87.37 (0.38)	12.22 (0.31)	5.34 (0.31)
	b = 0.1	86.94 (0.65)	13.41 (0.39)	5.69 (0.57)
	b = 0.5	88.48 (0.52)	11.12 (0.5)	4.37 (0.35)
	b = 0.9	88.98 (0.33)	10.98 (0.28)	4.64 (0.29)
k = 10, a = 0.5	b = 0	84.44 (2.43)	15.85 (2.39)	6.73 (2.47)
	b = 0.05	79.64 (3.1)	22.02 (5.15)	10.28 (4.06)
	b = 0.1	79.88 (2.63)	21.09 (3.59)	10.25 (1.85)
	b = 0.5	84.44 (2.54)	14.33 (1.78)	6.52 (2.83)
	b = 0.9	81.06 (2.35)	20.53 (4.52)	8.68 (3.36)
k = 10, b = 0.9	a = 0.005	73.91 (3.01)	28.02 (5.66)	13.85 (4.82)
	a = 0.01	72.41 (4.86)	25.57 (5.78)	13.66 (7.01)
	a = 0.02	72.03 (1.79)	31.25 (7.25)	17.26 (6.56)
	a = 0.05	73.2 (3.33)	30.41 (6.2)	17.96 (6.04)
	a = 0.1	75.28 (1.98)	23.96 (4.76)	10.13 (4.25)
	a = 0.5	81.06 (2.35)	20.53 (4.52)	8.68 (3.36)
	a = 0.8	85.99 (0.73)	13.76 (0.75)	6 (0.83)
	a = 0.9	87.27 (0.41)	13.72 (0.41)	5.68 (0.32)
	a = 1.0	88.98 (0.33)	10.98 (0.28)	4.64 (0.29)
k = 10, b = 0.5	a = 0.005	71.45 (3.81)	21.14 (4.37)	11.5 (5.46)
	a = 0.01	74.73 (6.24)	25.24 (3.84)	8.28 (4.35)
	a = 0.02	73.59 (3.72)	29.47 (6.89)	17.52 (6.13)
	a = 0.05	74.17 (3.88)	20.26 (4.15)	5.79 (3.7)
	a = 0.1	72.43 (2.75)	25.77 (5.41)	13.42 (4.89)
	a = 0.5	84.44 (2.54)	14.33 (1.78)	6.52 (2.83)
	a = 0.8	87.26 (0.41)	11.76 (0.24)	4.62 (0.21)
	a = 0.9	86.85 (0.54)	12.54 (0.44)	5.25 (0.48)
	a = 1.0	88.48 (0.52)	11.12 (0.5)	4.37 (0.35)
a = 1, b = 0.9	k = 10	88.98 (0.33)	10.98 (0.28)	4.64 (0.29)
	k = 100	88.19 (0.19)	11.15 (0.23)	4.67 (0.17)
	k = 500	87.98 (0.62)	11.33 (0.35)	4.72 (0.55)

Table 3. Ablation on k-NN label smoothing's hyperparameters: a, b, and k for the SVHN dataset.

C. Ablation Study

In Table 3, we report SVHN results ablating k-NN label smoothing's hyperparameters: k, a, and b. We observe the following trends: with a fixed to 1, both accuracy and churn improve with increasing b, and a similar relationship holds as a increases with b fixed to 0.9. Lastly, both key metrics are stable with respect to k.

D. Hyperparameter Search

Our experiments involved performing a grid search over hyperparameters. We detail the search ranges per method below.

k-NN label smoothing.

- $k \in [5, 10, 100, 500]$
- $a \in [0.005, 0.01, 0.02, 0.05, 0.1, 0.5, 0.8, 0.9, 1.0]$
- $b \in [0, 0.05, 0.1, 0.5, 0.9]$

Anchor.

- $a \in [0.005, 0.01, 0.02, 0.05, 0.1, 0.5, 0.8, 0.9, 1.0]$

ℓ_1, ℓ_2 Regularization.

- $a \in [0.001, 0.01, 0.05, 0.1, 0.2, 0.5]$

Co-distill

- $a \in [0.001, 0.01, 0.05, 0.1, 0.2, 0.5]$
- $n_{\text{warm}} \in [1000, 2000]$

Bi-tempered

- $t_1 \in [0.3, 0.5, 0.7, 0.9]$
- $t_2 \in [1., 2., 3., 4.]$
- n_{iters} always set to 5.

Mixup

- $a \in [0.2, 0.3, 0.4, 0.5]$

Ensemble

- $m \in [3, 5]$