Appendix

A Pseudocode for A-GBM

Algorithm 3 has the pseudocode for the AGBM procedure introduced in Section 2. The GBFS training procedure is identical, except with the function being optimized being un-normalized.

Algorithm 3 Pseudocode for A-GBM

Require: Data $\{x_i, y_i\}$, $i = 1, .., n$, shrinkage ϵ, iterations N, penalty parameter μ, tree growth parameter α

1: model $H = 0$, residues $g_i = y_i$, $i = 1, 2, .., n$. and selected feature set $\Omega = \emptyset$
2: **for** $k = 1, 2, \ldots N$ **do**
3: Fit a tree h_k using μ to minimize (1) in every split and α as stopping criteria
4: $H = H + \epsilon h_k$
5: $g_i = y_i - H(x_i)$
6: $\Omega = \Omega \cup \{j,$ tree $h_k$uses feature $f_j\}$
7: **end for**
8: Output H and Ω

B Theoretical analysis of GTGBM

B.1 Notations and Setup

Consider $(\boldsymbol{X}, Y) \sim \mathbb{P}$. Y is the label and we have d features: $\boldsymbol{X} = (X_1, .., X_d)$. $X_1, , ., X_d$ are independent with each other (not assuming have the same distribution) and $0 \leq X_i \leq 1$ (as GTGBM first standardizes the feature value to be within $[0, 1]$). Assume there is an unknown subset $S^* \subset [d]$,$|S^*| = s$, such that

$$Y = \mu + \sum_{i \in S^*} f_i(X_i) + \epsilon, \tag{7}$$

where $\mu = \mathbb{E}Y$ is the population mean and ϵ is noise that has mean 0 and is independent with $\boldsymbol{X}$. f_is are unknown univariate functions. To make the model identifiable, we can assume without loss of generality that

$$\mathbb{E}f_i(X_i) = 0, i \in S^* \tag{8}$$

This is called a sparse additive model. For the setups of GTGBM, we independently generated $p = \left\lceil es\log(\frac{s}{\delta})\right\rceil$ random subsets of $[d]$: $S_1, .., S_p$, where $e = 2.71828..$ is the base of natural logarithm and $\delta \in (0, 1)$. From Theorem 4.1, with high probability ($\geq 1 - \delta$), for every relevant features (X_i,$i \in S^*$), there is a random subset that exactly covers this feature.

B.2 Proof of Theorem 4.1

Proof. Suppose we have d features, and without loss of generality the active features are $f_1, \ldots, f_s \in \{1, 2, \ldots, d\}$. We generate iid subsets $S_1, \ldots, S_p \subset [d]$, such that $\forall j \in [d], \quad P(j \in S_i) = 1/s$. We want to show that the probability that exactly one of the relevant features lies in one of the random groups we create is larger than $1 - \delta$. We do this by obtaining an upper bound on it's complement. For convenience, we use the following shorthands: $\{f_1, \ldots, f_s\} := \Omega, \quad \{S_1, \ldots, S_p\} := \mathcal{S}$. We bound the probability of the complement of the event we are interested as follows:

$$\begin{aligned}
&P(\exists j \in \Omega \quad : \forall S \in \mathcal{S}, j \notin S \ \textbf{ OR } \ \exists j' \neq j : j' \in S, \quad j' \in \Omega) \\
&\leq s(1 - P(f_1 \in S_1 \ \textbf{ and } \ \forall j' \neq f_1, \quad j' \in \Omega, \quad j' \notin S_1))^p \\
&= s\left(1 - \frac{1}{s}\left(1 - \frac{1}{s}\right)^{s-1}\right)^p \\
&\leq s\exp\left(-\frac{p}{s}\left(1 - \frac{1}{s}\right)^{s-1}\right) \\
&\leq s\exp\left(-\frac{p}{es}\right) \leq \delta
\end{aligned} \tag{9}$$

Where the first inequality follows from the union bound, the second inequality follows from Bernoulli's inequality. The final inequality in (9) holds so long as p satisfies the condition in the statement of the Theorem. □

B.3 Theoretical split criterion in GTGBM

A key component of tree algorithms are the rules for splitting a node. For the classical CART algorithm, we greedily build the tree by splitting with a feature and a threshold such that in the child nodes the sample are most homogeneous measured by square error loss. Mathematically, the population version of the split criterion can be written as a function $L(Z, t)$ of split feature Z (including the "peusdo" feature created by GT-GBM) and threshold $t \in \mathbb{R}$:

$$\begin{aligned}
L(Z,t) \quad = \quad & \mathbb{E}[(Y - \mathbb{E}\left[Y|Z < t\right])^2 1\{Z < t\} \\
& + (Y - \mathbb{E}\left[Y|Z \geq t\right])^2 1\{Z \geq t\}]
\end{aligned} \tag{10}$$

Note that the split function is invariant with a shift of a constant in Y, so we may assume $\mu = \mathbb{E}Y = 0$ without loss of generality. Then some calculations lead to

$$L(Z,t) = \mathbb{E}Y^2 - \frac{\mathbb{E}^2\left[Y1\{Z < t\}\right]}{\mathbb{P}\left(Z < t\right)} - \frac{\mathbb{E}^2\left[Y1\{Z \geq t\}\right]}{\mathbb{P}\left(Z \geq t\right)} \tag{11}$$

Since $\mathbb{E}Y = 0$, we have

$$\mathbb{E}\left[Y1\{Z < t\}\right] = -\mathbb{E}\left[Y1\{Z \geq t\}\right].$$

Let $M(Z,t) = \mathbb{E}\left[Y1\{Z \geq t\}\right]$, we can further write

$$L(Z,t) = \mathbb{E}Y^2 - \frac{M^2(Z,t)}{\mathbb{P}\left(Z<t\right)\mathbb{P}\left(Z\geq t\right)} \tag{12}$$

In the algorithm, we will choose (Z,t) that minimize $L(Z,t)$ (the sample estimated version, see next section) which is equivalent to maximize $\frac{M^2(Z,t)}{\mathbb{P}(Z<t)\mathbb{P}(Z\geq t)}$. Note that if Z and Y are independent, then

$$\begin{aligned} M(Z,t) &= \mathbb{E}\left[Y1\{Z \geq t\}\right] \\ &= \mathbb{E}\left[Y\right]\mathbb{P}\left(Z\geq t\right) \\ &= 0 \end{aligned} \tag{13}$$

Thus no variance reduction takes into place. Let's recall the GTGBM procedure to find the split feature: for the p independently generated random group of features, we perform binary search. That is, for random subset $S \subset [d]$,write

$$Z_S = \sum_{i\in S} X_i,$$

we split S into left-half S_L and right-half S_R and calculate $\inf_t L(Z_{S_L},t)$ and $\inf_t L(Z_{S_R},t)$. We select the half with smaller value and recursively find the candidate split feature. We find the candidate split features for all p random subsets of features, and we choose the best split feature among them. Now we show that, if we have access to the theoretical split criterion (that corresponds to the ideal situation that we have infinite amount of data), the GTGBM split-finding procedure can actually find the best split feature. We only need to show that all relevant features: X_i,$i \in S^*$ are among the candidate split features. For $i \in S^*$, from Theorem 4.1 we know that there is a random subset $S \in \{S_{1,..},S_p\}$ such that $i \in S$ and for any $i' \in S^*, i' \neq i$, we have $i' \notin S$. Now we show that when we perform binary search on S, the half that contains the important feature index i is always been selected. Thus the output of binary search on S is exactly this index i. Suppose the left half S_L contains i. Then S_R doesn't contain i and also doesn't contain any $i' \in S^*, i' \neq i$ since S doesn't contain them. Thus Z_{S_R} is independent with Y, so $M(Z_{S_R},t) = 0$ for any t. On the other hand

$$\begin{aligned} M(Z_{S_L},t) &= \mathbb{E}\left[Y1\{Z_{S_L} \geq t\}\right] \\ &= \sum_{i\in S^*} \mathbb{E}\left[f_i(X_i)1\{Z_{S_L} \geq t\}\right] \\ &= \mathbb{E}\left[f_i(X_i)1\{X_i + \sum_{i'\neq i, i'\in S_L} X_{i'} \geq t\}\right] \end{aligned} \tag{14}$$

We can choose t such that $M(Z_{S_L},t) \neq 0$, as long as f_i is not degenerated. Thus we always have

$$\begin{aligned} \inf_t L(Z_{S_L},t) &\leq L(Z_{S_L},t) \\ &= \mathbb{E}Y^2 - \frac{M^2(Z_{S_L},t)}{\mathbb{P}\left(Z_{S_L}<t\right)\mathbb{P}\left(Z_{S_L}\geq t\right)} \\ &< \mathbb{E}Y^2 \\ &= \inf_t L(Z_{S_R},t). \end{aligned}$$

But in reality, we are using sample version of split function that only approximates the theoretical split function. So the condition for GTGBM to successfully find the best split feature depends on how the approximation error between theoretical split function and empirical split function and magnitude of $M^2(Z_{S_L},t)$ (still assumes S_L is the half that contains the relevant feature index) change with sample size n at a node and total number of features d. Intuitively, the increase of dimension d will harm the signal strength $M^2(Z_{S_L},t)$ since the irrelevant part $\sum_{i'\neq i, i'\in S_L} X_{i'}$ in equation (14)becomes more dominant. We rigorously showed that (see lemma B.1), under fairly general condition we have

$$M^2(Z_{S_L},t) \gtrsim \frac{1}{|S_L|} \gtrsim \frac{s}{d}. \tag{15}$$

Then we just need to know how well we can approximate theoretical split function by the empirical ones with sample size n.

B.4 Empirical split criterion in GTGBM

Suppose we have $i.i.d$ sample in a node $(\boldsymbol{X}_i, Y_i) \sim \mathbb{P}, i = 1,2,..,n$. $\boldsymbol{X}_i = (X_{i1},..,X_{id})$. The empirical split function is

$$L_n(Z,t) = \frac{1}{n}\left(\sum_{i:Z_i<t}\left(Y_i - \bar{Y}_L\right)^2 + \sum_{i:Z_i\geq t}\left(Y_i - \bar{Y}_R\right)^2\right) \tag{16}$$

where $\bar{Y}_L = \frac{\sum_i Y_i 1\{Z_i<t\}}{\sum_i 1\{Z_i<t\}}$, $\bar{Y}_R = \frac{\sum_i Y_i 1\{Z_i\geq t\}}{\sum_i 1\{Z_i\geq t\}}$ and Z_i, $i = 1,..,n$ is the $i.i.d$ sample for split feature Z. With a standard argument and concentration inequality (see lemma B.2), we can prove

$$\sup_t |L_n(Z,t) - L(Z,t)| = O_p(\frac{1}{\sqrt{n}}). \tag{17}$$

Thus with high probability, we have

$$\begin{aligned} \inf_t L_n(Z_{S_L},t) &\leq \inf_t L(Z_{S_L},t) + O(\frac{1}{\sqrt{n}}) \\ &\lesssim \mathbb{E}Y^2 - \frac{s}{d} + O(\frac{1}{\sqrt{n}}) \\ &= \inf_t L(Z_{S_R},t) - \frac{s}{d} + O(\frac{1}{\sqrt{n}}) \\ &\leq \inf_t L_n(Z_{S_R},t) - \frac{s}{d} + O(\frac{1}{\sqrt{n}}) \end{aligned} \tag{18}$$

The first and last inequality is from (17) and the second inequality is from (15). So, we only need $n \gtrsim (\frac{d}{s})^2$ for GTGBM to find the best split variables.

B.5 Proof of Theorem 4.2

The above subsections did some intuitive calculations that motivate the claim. This subsection aims at providing rigorous statement and filling the gaps. First let's recall the conditions assumed in theorem 4.2.

Assume

1. *X_i has bounded probability density function $p_i(x)$ and positive variance. Denote $B_d^2 = Var(X_1 + .. + X_d)$. Suppose $B_d \to \infty, d \to \infty$ and the limit*
$$\eta = \lim_{d\to\infty} \frac{\mu_d}{B_d}$$
exists, where $\mu_d = \mathbb{E}[X_1 + .. + X_d]$.
2. *The unknown functions in (7) are bounded monotone functions.*

We have following two lemmas:

Lemma B.1. *Recall the notation, for subset $S \subset [d]$, $Z_S = \sum_{i\in S} X_i$. if there is an index $i \in S^*$ that $i \in S$ and for any $i' \neq i, i' \in S^*$ we have $i' \notin S$. Assume the unknown function component f_i is bounded monotone. Also assume condition 1 in theorem 1. Then there exists constants t_0 , $d_0 > 0, c_0 > 0$ that only depend on the unknown functions in (7) and η, such that when $|S| \geq d_0$, we have*

$$L(Z_S, t_0) \leq \mathbb{E}Y^2 - \frac{c_0}{|S|} \tag{19}$$

proof of lemma B.1. From (13), we only need to show that there exists constants t_0 , $d_0 > 0, c_0 > 0$, such that

$$\frac{M^2(Z_S, t_0)}{\mathbb{P}(Z_S < t_0)\,\mathbb{P}(Z_S \geq t_0)} \geq \frac{c_0}{|S|} \tag{20}$$

First let's look at the numerator. From (14), we have

$$M(Z_S, t) = \mathbb{E}\left[f_i(X_i)1\{X_i + \sum_{i'\neq i, i'\in S} X_{i'} \geq t\}\right] \tag{21}$$

Denote $S' = S\backslash\{i\}$ and X_i and $Z_{S'}$'s probability density function as $p_i(x)$ and $p_{Z_{S'}}(z)$ respectively. Since X_i and $Z_{S'}$ are independent, we have

$$\begin{aligned}
&\mathbb{E}[f_i(X_i)1\{X_i + Z_{S'} \geq t\}] \\
&= \int_{x+z\geq t} f_i(x)p_i(x)p_{Z_{S'}}(z)dzdx \\
&= \int f_i(x)p_i(x) \int_{z\geq t-x} p_{Z_{S'}}(z)dzdx
\end{aligned}$$

On the other hand, since f_i is monotone function (without loss of generality assume it's monotone increasing), then there exists $t_0 \in [0, 1]$ such that $f_i(t_0) = 0$ and $f_i(t) > 0$ for $t > t_0$ and $f_i(t) < 0$ for $t < t_0$. Then, $\mathbb{E}[f_i(X_i)1\{X_i + Z_{S'} \geq t_0\}]$ can be written as

$$\begin{aligned}
&\int_{x\geq t_0} f_i(x)p_i(x) \int_{z\geq t_0-x} p_{Z_{S'}}(z)dzdx \\
&\quad + \int_{x<t_0} f_i(x)p_i(x) \int_{z\geq t_0-x} p_{Z_{S'}}(z)dzdx \\
&= \int_{1\geq x\geq t_0} f_i(x)p_i(x) \int_{t_0-x}^{0} p_{Z_{S'}}(z)dzdx \\
&\quad - \int_{0\leq x<t_0} f_i(x)p_i(x) \int_{0}^{t_0-x} p_{Z_{S'}}(z)dzdx
\end{aligned} \tag{22}$$

The equation is from the fact that $\int_{x\geq t_0} f_i(x)p_i(x)dx + \int_{x<t_0} f_i(x)p_i(x)dx = \mathbb{E}[f_i(X_i)] = 0$. Let $m_{Z_{S'}} = \min_{z\in[t_0-1,t_0]} p_{Z_{S'}}(z)$. Then the right hand side of (22) is lower bounded by

$$m_{Z_{S'}} \int_0^1 (x - t_0)f_i(x)p_i(x)dx. \tag{23}$$

Note that $(x - t_0)f_i(x)p_i(x) \geq 0$ for any $x \in [0, 1]$ and there exists a positive measure set such that $(x - t_0)f_i(x)p_i(x) > 0$ (otherwise X_i is degenerated). Thus we denote $v_0 = \int_0^1 (x - t_0)f_i(x)p_i(x)dx$ and $v_0 > 0$. Now let's look at the other factor $m_{Z_{S'}}$ in (23) Denote $\tilde{Z}_{S'} = \frac{Z_{S'} - \mathbb{E}Z_{S'}}{\sqrt{\mathrm{Var}(Z_{S'})}}$ as standardized $Z_{S'}$, then we have

$$p_{Z_{S'}}(z) = \frac{1}{\sqrt{\mathrm{Var}(Z_{S'})}} p_{\tilde{Z}_{S'}}\left(\frac{z - \mathbb{E}Z_{S'}}{\sqrt{\mathrm{Var}(Z_{S'})}}\right). \tag{24}$$

From condition 1 and the well known local limit theorem, the standardized density function $p_{\tilde{Z}_{S'}}(z)$ uniformly converge to standardized normal density $\phi(z)$ as $|S'| \to \infty$. Moreover

$$\lim_{|S'|\to\infty} \sqrt{\mathrm{Var}(Z_{S'})} m_{Z_{S'}} = \phi(-\eta) \tag{25}$$

since from condition 1, we have $\lim_{|S'|\to\infty} \frac{z - \mathbb{E}Z_{S'}}{\sqrt{\mathrm{Var}(Z_{S'})}} = -\eta, \forall z \in [t_0 - 1, t_0]$. Combined with (21)(22)(23), we conclude that there exists a constant d_1 such that when $|S| > d_1$, we have

$$M^2(Z_S, t) \geq \frac{v_0^2\phi^2(-\eta)}{2\mathrm{Var}(Z_{S'})} \geq \frac{v_0^2\phi^2(-\eta)}{2|S|} \tag{26}$$

where the second inequality follows from $\mathrm{Var}(Z_{S'}) = \sum_{i\in S'} \mathrm{Var}(X_i) \leq |S'| < |S|$ since $X_i \leq 1$. For the denominator in (20), from Central Limit Theorem, we have

$$\mathbb{P}(Z_S < t_0) = \mathbb{P}\left(\frac{Z_S - \mathbb{E}Z_S}{\sqrt{\mathrm{Var}(Z_S)}} < \frac{t_0 - \mathbb{E}Z_S}{\sqrt{\mathrm{Var}(Z_S)}}\right) \to \Phi(-\eta)$$

as $|S| \to \infty$, where Φ is the distribution function of standard normal. Thus there exists a constant d_2, such that when $|S| > d_2$, we have

$$\mathbb{P}\left(Z_S < t_0\right)\mathbb{P}\left(Z_S \geq t_0\right) \leq 2\Phi(-\eta)\Phi(\eta). \tag{27}$$

Thus combine (26)(27), we showed that for $|S| > d_0 = \max\{d_1, d_2\}$, we have

$$\frac{M^2(Z_S, t_0)}{\mathbb{P}\left(Z_S < t_0\right)\mathbb{P}\left(Z_S \geq t_0\right)} \geq \frac{c_0}{|S|}$$

where $c_0 = \frac{v_0^2\phi^2(-\eta)}{4\Phi(\eta)\Phi(-\eta)} > 0$. That concludes the proof. □

Lemma B.2. *There exists positive constants c_1, c_2 that only depend on the unknown fixed component functions such that for any $0 < x < 1$*

$$\mathbb{P}\left(\sup_t |L_n(Z,t) - L(Z,t)| \leq x\right) \geq 1 - c_1\exp(-c_2 n x^2) \tag{28}$$

proof of lemma B.2. Let $\mu_L = \frac{\mathbb{E}[Y1Z<t]}{\mathbb{P}(Z<t)}$, $\mu_R = \frac{\mathbb{E}[Y1Z\geq t]}{\mathbb{P}(Z\geq t)}$ and $n_L = \sum_i 1\{Z_i < t\}$, $n_R = \sum_i 1\{Z_i \geq t\}$. Define

$$\tilde{L}_n(Z,t) = \frac{1}{n}\left(\sum_{i:Z_i<t}(Y_i - \mu_L)^2 + \sum_{i:Z_i\geq t}(Y_i - \mu_R)^2\right) \tag{29}$$

Then

$$\begin{aligned}\tilde{L}_n(Z,t) - L_n(Z,t) = \\ \frac{1}{n}\sum_{i:Z_i<t}(\bar{Y}_L - \mu_L)(2Y_i - \mu_L - \bar{Y}_L) \\ + \frac{1}{n}\sum_{i:Z_i\geq t}(\bar{Y}_R - \mu_R)(2Y_i - \mu_R - \bar{Y}_R) \\ = \frac{n_L}{n}(\bar{Y}_L - \mu_L)^2 + \frac{n_R}{n}(\bar{Y}_R - \mu_R)^2 \end{aligned} \tag{30}$$

Also we can write $\bar{Y}_L - \mu_L$ as

$$\begin{aligned}\frac{1}{n_L}\sum_i (Y_i 1\{Z_i < t\} - \frac{\mathbb{E}[Y1\{Z<t\}]}{\mathbb{P}(Z<t)}) \\ = \frac{n}{n_L}\frac{1}{n}\sum_i (Y_i 1\{Z_i < t\} - \mathbb{E}[Y1\{Z<t\}]) \\ + \mathbb{E}[Y1\{Z<t\}](\frac{n}{n_L} - \frac{1}{\mathbb{P}(Z<t)})\end{aligned} \tag{31}$$

Since $1\{Z_i < t\} - \mathbb{P}(Z < t)$ and $Y_i 1\{Z_i < t\} - \mathbb{E}[Y1\{Z < t\}]$ are i.i.d mean 0 bounded random variables (and the bound doesn't depend on t), from Bernstain inequality, for any t and $x > 0$, we have

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_i 1\{Z_i < t\} - \mathbb{P}(Z<t)\right| \geq x\right) \leq 2\exp(-c_1 n x^2) \tag{32}$$

and

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_i Y_i 1\{Z_i < t\} - \mathbb{E}[Y1\{Z<t\}]\right| \geq x\right) \leq 2\exp(-c_2 n x^2) \tag{33}$$

where c_1, c_2 are positive constants that don't depend on t. Combine (31)(32)(33), with proper change of the constants c_1, c_2, we conclude that, for all t and any $x > 0$

$$\mathbb{P}\left(\left|\bar{Y}_L - \mu_L\right| \geq x\right) \leq c_1\exp(-c_2 n x^2) \tag{34}$$

We can apply the same argument to $\bar{Y}_R - \mu_R$. Thus for any $x > 0$,

$$\mathbb{P}\left(\sup_t \left|\tilde{L}_n(Z,t) - L_n(Z,t)\right| \geq x\right) \leq c_1\exp(-c_2 n x) \tag{35}$$

for proper constants c_1, c_2. When $x < 1$, the right hand side of (35) $\leq c_1\exp(-c_2 n x^2)$. Thus we only need to prove

$$\mathbb{P}\left(\sup_t \left|\tilde{L}_n(Z,t) - L(Z,t)\right| \geq x\right) \leq c_1\exp(-c_2 n x^2) \tag{36}$$

This also follows from Bernstain inequality, since $\tilde{L}_n(Z,t) - L(Z,t)$ is the average of i.i.d mean 0 random variables

$$w_i := (Y_i - \mu_L)^2 1\{Z_i < t\} + (Y_i - \mu_R)^2 1\{Z_i \geq t\} - L(Z,t)$$

w_i is also bounded (since Y_i are bounded) and the bound doesn't depend on t. □

Now let's go back to the proof of main theorem. From (18) and lemma B.1 and B.2, the failure probability of identifying the correct half group that contains the important feature is bounded by $c_1\exp(-c_2 n x^2)$ with $x = \frac{c_0 s}{4d}$. Given $\delta \in (0,1)$, since GTGBM performs at most $es\log(\frac{2s}{\delta})\log_2(\frac{d}{s})$ times of comparing two splitted groups of variables (assume we generate $es\log(\frac{2s}{\delta})$ random subsets) , by union bound and theorem 4.1, the overal failure probability is bounded by

$$\frac{\delta}{2} + c_1 es\log(\frac{2s}{\delta})\log_2(\frac{d}{s})\exp(-c_2 n x^2)$$

with $x = \frac{c_0 s}{4d}$. Solve n for

$$c_1 es\log(\frac{2s}{\delta})\log_2(\frac{d}{s})\exp(-c_2 n x^2) \leq \frac{\delta}{2}$$

with $x = \frac{c_0 s}{4d}$ gives the conclusion.

C Optimal Hyperparameters to Reproduce Results on Public Datasets

Here we give additional details required to reproduce the results we obtained on all 3 public datasets. We used the train/test split that was provided online in all the cases: 6000/1000 for Gisette, 80000, 20000 for Epsilon and $100K, 100K$ for Flight Delay

For tuning the hyperparameters, we further split the train set into an 80-20 train and validation set, and cross-validate on the latter. Table 5 lists the optimal hyperparameters for all the algorithms used. 'α' is the minimum fraction of data in an internal node (parameter that controls the size of a single tree).

Table 5: Optimal hyperparameters for all methods

Dataset	Method	μ	shrinkage ϵ	α
Gisette	GBDT	-	0.1	0.02
	GBFS	1.1	0.1	0.02
	A-GBM	0.01	0.1	0.02
	GT-GBM	0.001	0.1	0.02
Epsilon	GBDT	-	0.1	0.02
	GBFS	2.0	0.1	0.02
	A-GBM	0.0004	0.1	0.02
	GT-GBM	0.0001	0.1	0.02
Flight	GBDT	-	0.1	0.1
	GBFS	4	0.1	0.1
	A-GBM	0.0004	0.1	0.1
	GT-GBM	0.0002	0.1	0.1

D Performance When All Features Are Used

For the sake of completeness, we provide the optimum hyperparameter values as well as the results obtained on the public datasets when we use all the available features to train the model. Note that we report this performance for the sake of comparison, and as we mentioned earlier, such a method is not practical in the applications we consider. The results are provided in Table D

Table 6: Performance of the full GBDT model on all public datasets

Dataset	Method	shrinkage ϵ	α	AUC
Gisette	GBDT-Full	0.1	0.02	99.33
Epslion	GBDT-Full	0.1	0.02	92.34
Flight	GBDT-Full	0.1	0.1	71.74

E Performance on Internal Classification Datasets

For the internal classification dataset, we compute the area under the ROC curve, and the Precision at 2. The task in both cases is to identify items in response to query-item pairs that have been marked as "incorrect." We see from Table 7 that GBDT-topK methods are suboptimal, and GT-GBM matches or outperforms GBFS, while being vastly superior in terms of training time.

F Multitask Results on M2

There are 4 countries in total. Again, we hypothesize that there will be features that might be common across countries that we can use, and country specific features that depend on the items available, and vagaries of the languages spoken in those countries. We aim to see if combining information from various sources and training joint models helps to achieve better metrics as compared to training models individually. Figure 5 again shows that the multitask GTGBM outperforms the single task and traditional multitask counterparts.

Table 7: Comparison of various methods for the classification tasks (C1 and C2). In both cases, GBDT-topK is suboptimal, and GT-GBM narrowly outperforms GBFS. Bold numbers indicate the best result.

Dataset	Measure	GBDT-topK	GBFS	GT-GBM
C1	AUC_ROC	0.918	**0.922**	0.920
	prec@k=2	0.751	0.770	**0.773**
	RMSE	0.260	**0.258**	**0.258**
C2	AUC_ROC	0.910	0.910	**0.912**
	prec@k=2	0.874	0.875	**0.878**
	RMSE	0.219	**0.218**	**0.218**

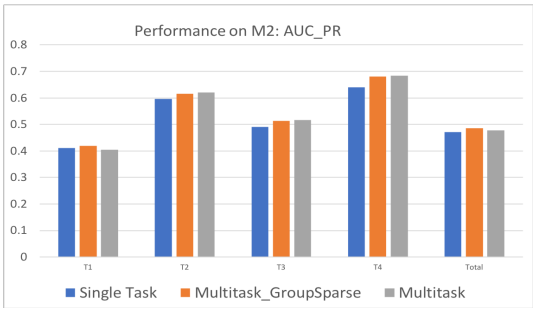

Figure 5: Performance on M2, for Area Under Precision-Recall curves. As in the previous experiment, using both task-specific and across-task features is beneficial. The performance boosts for tasks T2-T4 arise from using the data from T1, which has the largest and cleanest dataset.