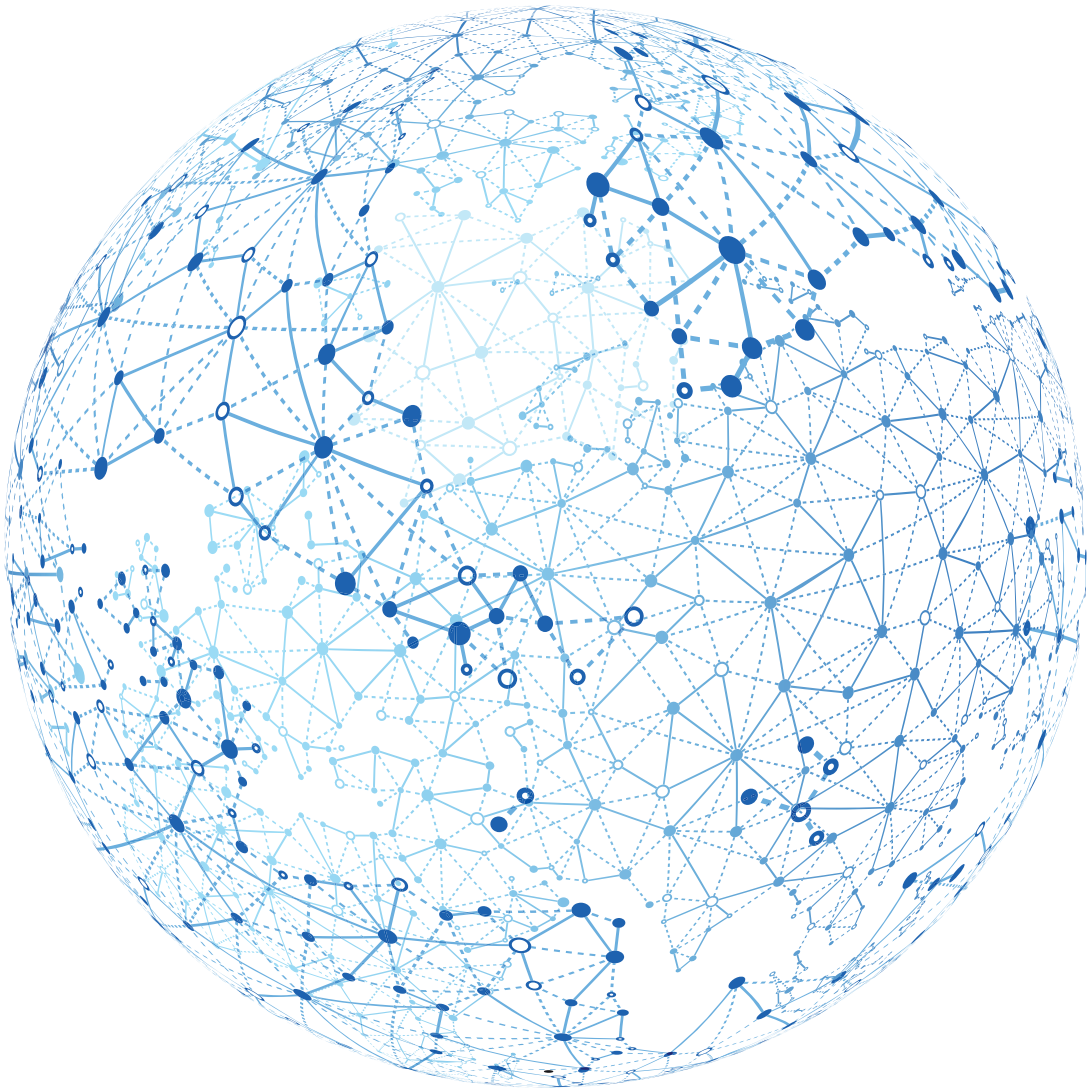
Deloitte.



Uncovering a new path for machine learning data classification

A Deloitte AI Institute publication

About the Deloitte AI Institute

The Deloitte AI Institute helps organizations connect all the different dimensions of the robust, highly dynamic and rapidly evolving AI ecosystem. The AI Institute leads conversations on applied AI innovation across industries, with cutting-edge insights, to promote human-machine collaboration in the "Age of With".

Deloitte AI Institute aims to promote the dialogue and development of artificial intelligence, stimulate innovation, and examine challenges to AI implementation and ways to address them. The AI Institute collaborates with an ecosystem composed of academic research groups, start-ups, entrepreneurs, innovators, mature AI product leaders, and AI visionaries, to explore key areas of artificial intelligence including risks, policies, ethics, future of work and talent, and applied AI use cases. Combined with Deloitte's deep knowledge and experience in artificial intelligence applications, the Institute helps make sense of this complex ecosystem, and as a result, deliver impactful perspectives to help organizations succeed by making informed AI decisions.

No matter what stage of the AI journey you're in; whether you're a board member or a C-Suite leader driving strategy for your organization, or a hands on data scientist, bringing an AI strategy to life, the Deloitte AI institute can help you learn more about how enterprises across the world are leveraging AI for a competitive advantage. Visit us at the Deloitte AI Institute for a full body of our work, subscribe to our podcasts and newsletter, and join us at our meet ups and live events. Let's explore the future of AI together.

www.deloitte.com/us/AIInstitute

Applying labels manually to data can be expensive and time consuming

One of the main problems that data scientists face today is unlabeled data. The algorithms that they use must have an initial amount of data labeled to learn upon. There are different routes to get there, but uncovering the best, most efficient and accurate route, needed further exploration. Deloitte's Innovation and Platforms Machine Learning research team took on the challenge.

Labels serve as the ground truth and are vital for machine learning algorithms to work. When creating a machine learning model, the typical approach is to take tens if not hundreds of thousands of rows of data that already have labels and feed them into a model. If done well, the data teaches it to learn the relevant patterns and solve the business case at hand.

If data is not already labeled, subject matter experts (SMEs) must spend many hours labeling enough data to test the model. This technique is referred to as passive learning, but it's not perfect. SMEs will often provide more labels in specific data subsets than are needed, while neglecting other subsets, which can heavily affect how the model performs.

Active learning is an alternative to this time-consuming and often biased approach to dealing with unlabeled data. Active learning takes a smaller amount of labeled data, runs it through a semi-supervised model iteratively, and uses these iterations to select the most useful new rows of data to label. The technique offers a faster alternative to creating labeled data while also providing more beneficial data for the model. Why?

The approach results in a representative sample of training data with typically very little bias because the machine is choosing what's most relevant instead of a human.

To compare the two approaches, both types of learning were tested. The active learning model yielded comparable results to the passive learning model with as little as a tenth of the labeled data. The performance of the active learning model was also 8 percent more accurate on average in classifying data than the passive learning approach when the model was allowed to continue learning past the number of rows from the passive approach.

Introduction

Business problem: Insufficient labeled data to adequately train a model

The problem that prompted research into active learning was having a lack of training data to reflect a business's spending. The datasets that a data science team typically receives already have labeled and unlabeled data. This often leads to using the labeled data for training, test and validation sets.

In a recent problem, completely unlabeled transactional data was received. Data scientists needed to identify which transactions should be used to train the model, but the SME's time was limited. A small subset of training data would need to be selected, rather than labeling a large dataset.

The goal of the research was to find a cost effective and quick method to label data from this unlabeled dataset and then create a model that performed in a comparable manner to past models. The data was high-level accounting information that specifically covered accounts payable data. There was also unstructured invoice text data, which included: the state where the transaction took place, accounting category, vendor name, invoice description and total amount as well as other text data that had been scanned directly from the invoice using Optical Character Recognition (OCR).

Passive learning

Passive learning is the typical approach for training a machine learning model and it applies a multi-step process to get training data, requiring an intense amount of work from SMEs. To identify subsets of data, clusters of transactions are created using the Birch clustering method. The Birch approach builds each cluster by gradually grouping transactions that are close to each existing cluster and measuring how distinct each cluster is [1]. For example, similar transactions can have different classifications based on what jurisdiction it originated in. A transaction for computers for administration purposes could be tax exempt in one jurisdiction, but fully taxed in another.

Next, the team spends two to three weeks labeling roughly 1,000 transactions that are sampled from each cluster. During labeling, they apply their understanding of the client's business situation and industry experience to identify the most appropriate label for each transaction. The team then uses the newly labeled transactions to train a model specifically for the client.

After the data is finished running through the model, the team verifies the performance on the test set. The model is used to predict further transaction classifications. However, in this scenario, a human makes all the choices about which data will be used to train the model.

Advantages

- Approach is convenient when labeled data already exists.
- A usable model can be created very quickly.
- Usually inexpensive as labeled data already exists.

Disadvantages

- Labeled dataset is likely to have biases that will influence the model.
- Important subsets of the data could be left unlabeled.

The basic approach to passive learning

Active learning

Deloitte researched active learning as an alternative to the current passive learning process in an effort to improve data classification methods. The hypothesis was that active learning could be used to better select transactions compared to the passive approach and it could do so with less strain on the SME.

Active learning is a special case of semi-supervised machine learning in which a learning algorithm can interactively query the user (or some other information source) to obtain the desired labels of new data points. In statistics, it is sometimes called optimal experimental design. This interaction between the model and the user or data source often occurs iteratively, with each iteration selecting new data points. It's not a linear process. It repeats several times.

There are similarities between active learning and another type of iteration-based learning called reinforcement learning. In both of these types of learning, the model can take action and adjust to select a better action or data point. In reinforcement learning, the model typically tries to maximize a measure of performance, such as points scored in a game, and there isn't any involvement from a human. Reinforcement learning simply tries to find the optimal actions needed to maximize this measure, sometimes leading to actions that a human would not have taken.

In active learning, the model is attempting to select the unlabeled data that will be the most informational.

The goal is to minimize the number of iterations and total number of labeled data points, while maximizing the accuracy of the predictive model.

In comparison to passive machine learning, where humans typically use existing labeled data to train the model, active learning reduces human involvement in the training-data selection process to a semi-supervisory role. Humans train the model with a small set of initial labeled data, then the model selects additional unlabeled data points that could provide the most information for the next training iteration. Over many iterations, the model will have selected the best set of data from which to train itself.

While active learning has many benefits, it can also have disadvantages if improperly used. If a dataset has a high degree of noise, the model is more likely to select data points for labeling based on this meaningless noise.

Humans train the model with a small set of initial labeled data, then the model selects additional unlabeled data points that could provide the most information for the next training iteration.

This disadvantage could lead the model to learn incorrect patterns because it sees the noise as significant.

The importance of experienced and knowledgeable SMEs is also greater when using active learning methods. If annotators incorrectly label the data they are given, the model will be more heavily influenced by these errors compared to the passive learning method. With passive learning, there is more than enough data to ignore one-off errors. Because active learning uses a smaller set of data, labeling errors have a more severe impact.

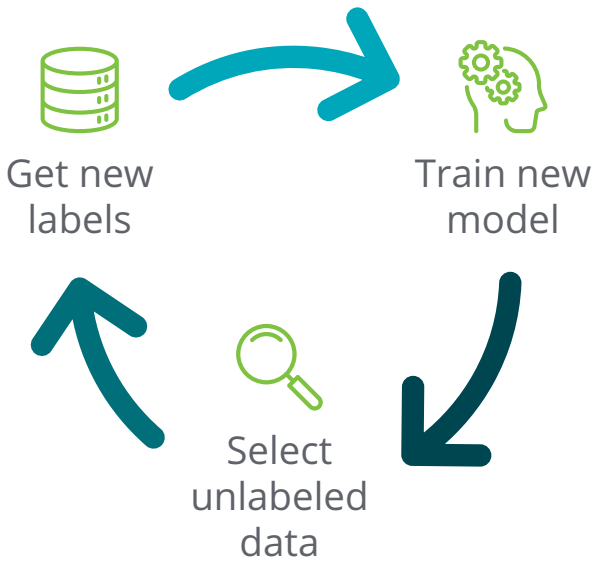
Advantages

- The entire dataset does not need to be labeled.
- A more representative set of data is selected.

Disadvantages

- The model is more likely to be sensitive to noise.
- The quality of SMEs' labeling can drastically affect performance.

The basic approach to active learning



Methodology

In active learning, there are several methods commonly used to select data points for labeling. Deloitte evaluated three methods of selecting data for further labeling.

Uncertainty sampling

Perhaps the simplest and most used framework is uncertainty sampling. In this type of sampling, the model selects the data point for which it has the least confidence in the label it has predicted [2]. This approach is often straightforward for probability-based learning models.

Margin sampling

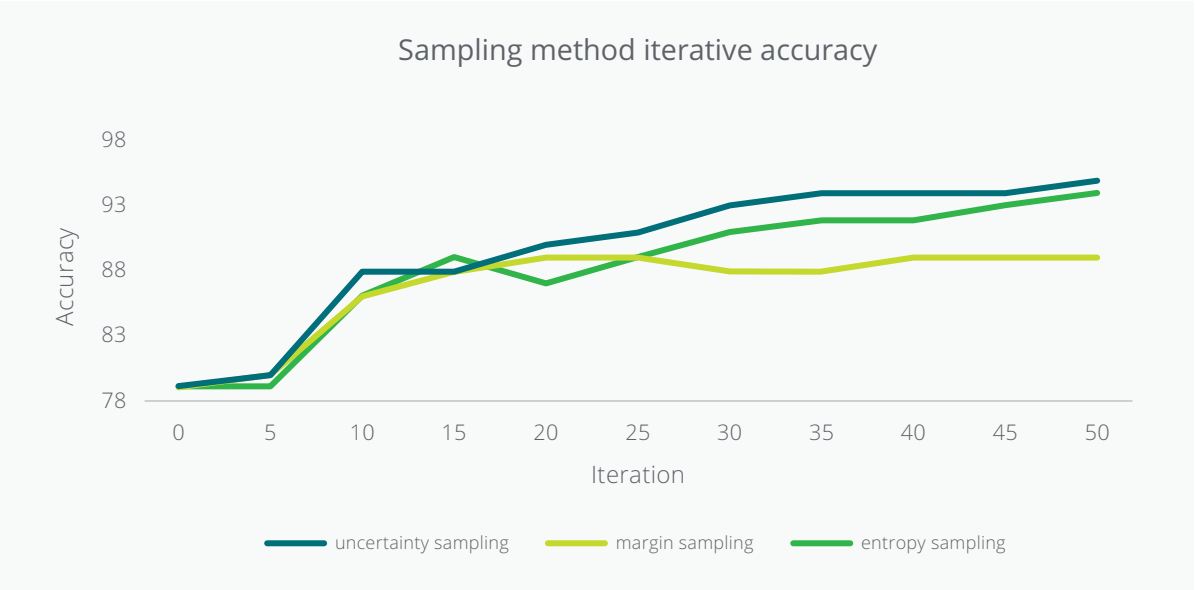
In margin sampling, the model selects the instance that has the smallest difference between the first and second most probable labels [3]. It relies on a basic machine learning model that attempts to separate groups of data to identify which transactions to sample, though it doesn't account for the distribution of the data. This can lead to oversampling a dense subset of the population [4].

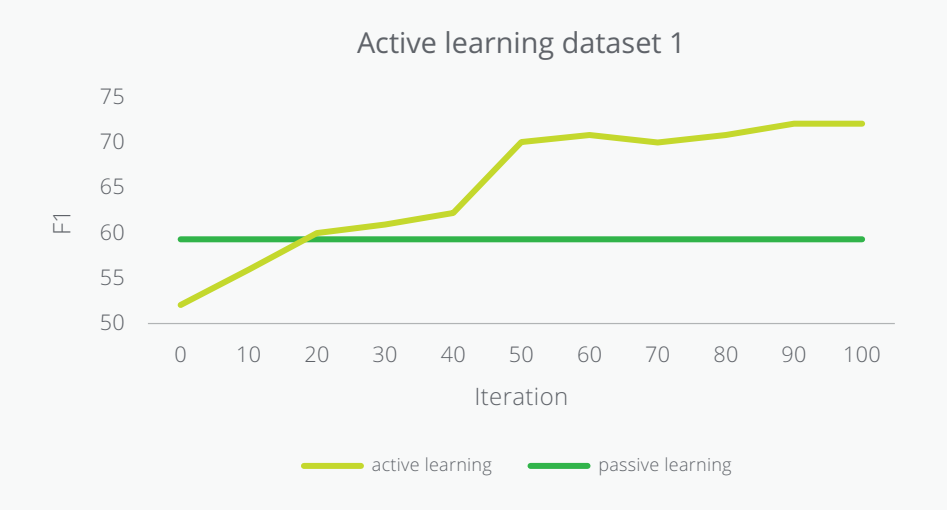
Entropy sampling

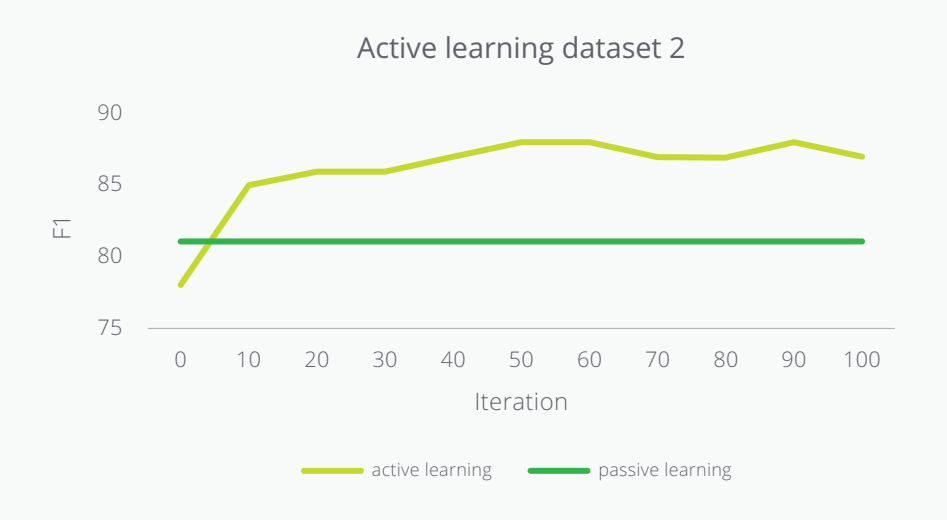
Entropy is an inquiry taken from information theory that measures disagreement among the likely labels for each data point. The entropy formula is applied to each instance and the data with the largest entropy is chosen as part of the next sample for labeling.

Choosing the sampling method

After initial tests, margin sampling was chosen for in-depth testing. While it wasn't consistently the best-performing selection method, its performance stabilized the fastest, which was more important for faster iterative testing. On multiple datasets, the three sampling methods resulted in similar test-model performance.



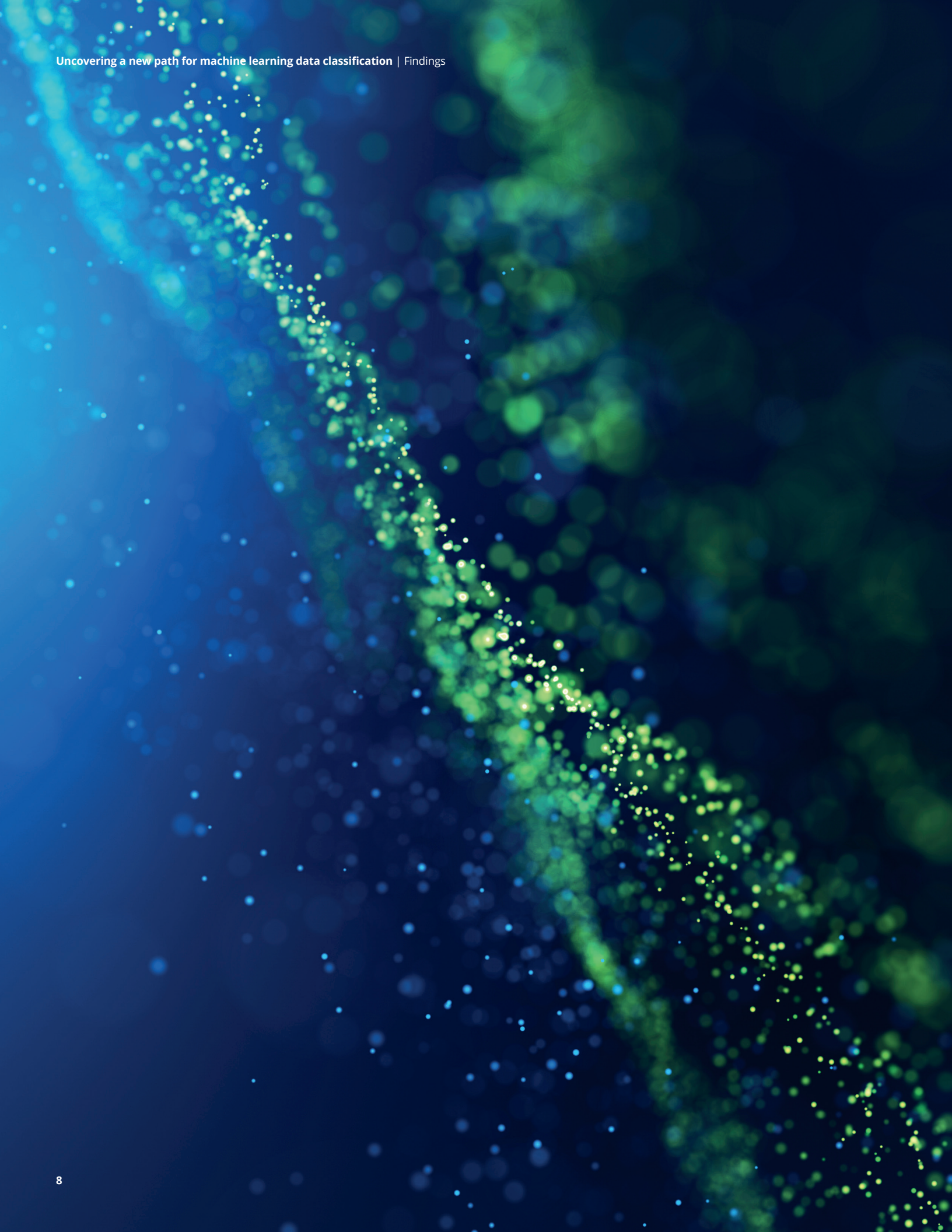




Modeling approach

Before any steps were taken, the dataset was rebalanced so classes were equally represented. Once this was complete, the process went through the following steps.

01. **The initial model was trained** using 5 percent of the labeled data and tested against 25 percent of the test data. This was iteration 0.
02. **The model predicted classifications for unlabeled data.**
03. **The model selected the most ambiguous predictions for ground-truth labeling.**
04. **SMEs labeled the selected data.**
05. **The model was trained with the additional labeled** instances and tested against the test data to measure performance.
06. **The process was repeated until the model's performance stabilized.** Several variables were investigated that could have affected the model's performance: the value distribution in selected dataset, the size of the entire dataset, the initial training batch size, and iterative labeling batch size.



Findings

The active learning model reached the same performance threshold as the traditional passive learning model, with only 26 percent of the labeled transactions of the traditional approach. The active learning model achieved the same level of performance and accuracy with just 130 labeled transactions. The passive learning model needed 500 labeled transactions.

Selection of initial training set matters

The performance of the active learning method depended on the size and nature of initial training data. Performance gradually improved and remained consistent if the initial training data size was anywhere between 20 to 40 data points, and if the classes were balanced.

Smaller batch sizes need more iterations

In each iteration of active learning, a small set or batch of data was selected for labeling. The number of transactions in each batch did not lead to a significant change in the number of labeled points needed for the model to converge.

The smaller batch sizes needed more iterations for the model's performance to plateau. In *Impact of Batch Size on Stopping Active Learning for Text Classification*, Beatty, Kochis and Bloodgood found that smaller batch sizes were more efficient from a model-learning perspective, though they needed more iterations with human annotators [5].

The dataset size should be just right

If the dataset size was large, active learning sometimes took many iterations to converge. Possible causes included more variation in the data or more edge cases that needed labels. The team did not investigate this further.

During testing, Deloitte noted that the influence of meaningless noise in the data was more pronounced with a smaller dataset size. There were larger swings in the model's performance from batch to batch with a small dataset.

Curiously, Deloitte observed that after the model converged, if active learning was used to select yet more training samples, the model's performance degraded. The team did not investigate why this happened, but our hypothesis is that as the model begins fitting on these marginal-decision boundary points, the effect of insignificant variations in the data cause the model to overfit.

Active learning is accurate

When reviewing the data, the distribution of values selected by the active learning method mirrored the distribution of values found in the full dataset. This is a significant improvement compared to the manual labeling of data, which often had highly skewed representations of several key features. The skew in the manually labeled data was likely from a bias introduced by a SME's focus in specific areas of potential recovery they knew were worth prioritizing, project experience or other background attributes. With active learning, the model selects a more representative dataset for training, likely reducing bias in the process.

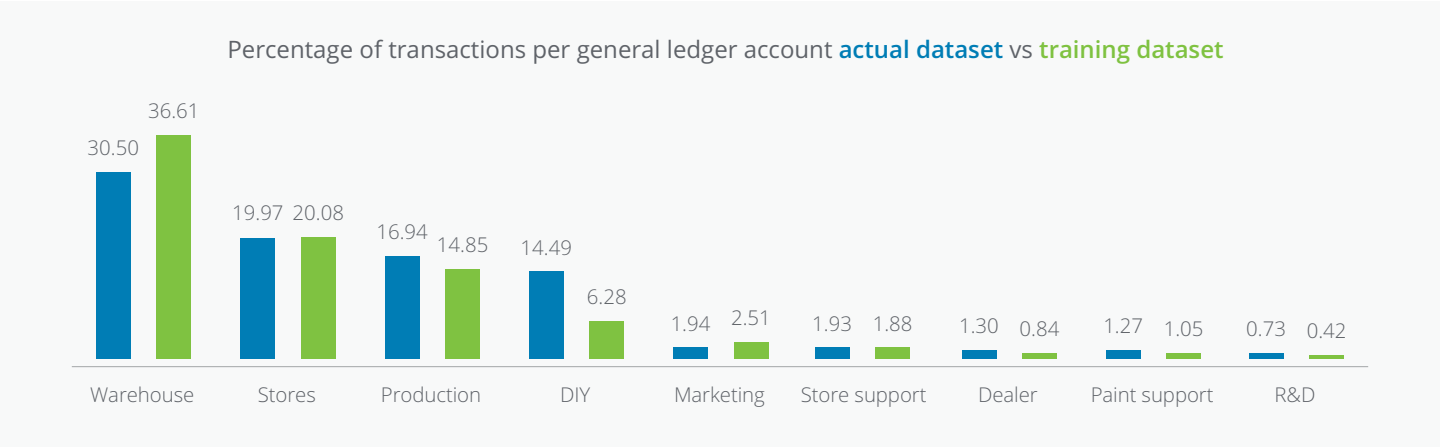
Stopping active learning

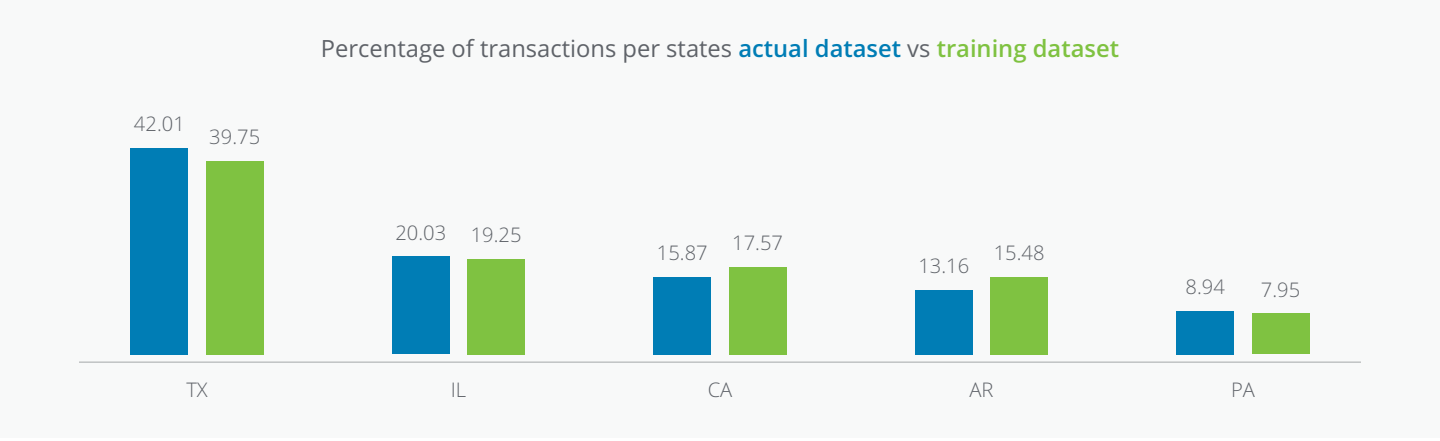
With active learning, it's important to define when to stop the iterations selecting new data:

1. There is no longer a return on investment for new labels. The model's performance has reached a level that is good enough for business needs or budget for SME labeling has run out.

2. A performance plateau is reached. In the testing conducted, the team continued to provide labeled data until a plateau in performance was reached. Past this point, adding more data wasn't helpful for the model at best, and at worst could lead to overfitting.

Distribution of values





However, if the cost of labeling is particularly expensive or time consuming, then it is likely that the labeling process would be stopped because of budgetary constraints before the model plateaus.

Challenges

Selecting a point at which to stop the active learning iterations was difficult as there was variation in the model's performance from batch to batch. There were many instances in our testing where the model's performance dropped for 2 to 3 iterations and then rose again to higher levels. As previously mentioned, allowing the model to continue training past the performance plateau generally led to overfitting to noise and reduced performance.

A potential solution to this challenge could be to save models at each iterative step and continue training the models until significant model degradation appears, allowing the most effective model version to be selected.

Another option would be to set the desired training sample size and to iterate sampling until that threshold is met [6]. Note that this method carries the risk that model performance hasn't reached optimization, but the method is appropriate when there are known high fixed costs with labeling, or when SME time is limited.

Other active learning use cases

Active learning can be a good option in situations where there is a large amount of unlabeled data, no existing labeled data or when labeling the data is time-consuming or expensive. Because the training process is much faster, it reduces cost. And because the model is selecting the most uncertain data points, it is more likely to identify rare situations that would normally be missed.

Active learning differs from traditional machine learning as it is focused on labeling a small portion of the entire dataset. By using active learning, a SME can focus on labeling the data that is most important for the use case, which can decrease the running time and increase the usability of the data.

With active learning, the model selects a more representative dataset for training, likely reducing bias in the process.

Conclusion

Active learning proved to be an effective method for better selecting training samples from an unlabeled set of data. Compared to passive learning, our research found that active learning models perform six to 12 percent better than passive learning models. This performance was different for each dataset, implying that the distribution of data heavily impacted the effectiveness of active learning.

While an active learning model isn't appropriate or beneficial for all cases -- especially when an adequate set of labeled data already exists -- it is able to quickly identify transactions that would provide the most information to a model.

The active learning model selects data that is more representative of the whole population, leading to a model that has less bias and is better equipped to generalize information.

In cases where a good training dataset of labeled transactions already exists, active learning doesn't provide a meaningful benefit. The increased time to iteratively add more training data would add further cost to a project, while also not providing a significant amount of new information to the model. However, in some cases the new method may fast become the norm.

Start a conversation

To learn more about active learning data classification, please reach out to one of our data scientists:

Gert De Geyter
SNET Data Science Lead
gedegeyter@deloitte.com

Ben Hooper
Data Scientist
bhooper@deloitte.com

Endnotes

1. Zhang, Ramakrishnan, Livny. (1997) *BIRCH: A New Data Clustering Algorithm and Its Applications*
2. Lewis and Gale. (1994) *A Sequential Algorithm for Training Text Classifiers*
3. Balcan, Broder, and Zhang. (2007) *Margin Based Active Learning*
4. Zhou and Sun. (2014) *Improved Margin Sampling for Active Learning*
5. Beatty, Kochis, Bloodgood. (2018) *Impact of Batch Size on Stopping Active Learning for Text Classification*
6. Schein and Ungar. (2007) *Active learning for logistic regression: An evaluation*

Deloitte.