Beyond Yes or No: Making Reliable Decisions and improving personalized customer service

Akash Singh*[1], Ashwin Ittoo†[1], Pierre Ars‡[2], Francois Dehouck[2], Francois Collienne[3], Norman Marlie[3], Tom To Hoang[2], and Nicolas Dumazy[3]

[1]HEC, School of Management of the University of Liège
[2]Ethias Insurance, Belgium
[3]NRB Group

September 2024

Abstract

This white paper explores how uncertainty tools can be used to improve personalized customer service.

Uncertainty is inherent in any machine learning predictive model. There are no “perfect models”, partly due to the *curse of dimensionality* and the challenges of avoiding any biases and misclassifications.

We aim to demonstrate how an insurance company can benefit from the *uncertainty* of machine learning predictions in order to develop methods that allow for the allocation of an uncertainty parameter to the predictions provided for a given profile/customer x.

The benefits of scrutinizing *uncertainty* are numerous and often aligned with customer interests:

- It can help to appreciate the weak points of a predictive model and thus improve them.
- It enables the definition of the Next Best Action (NBA) with a “full understanding of the facts”.
- It facilitates the analysis of marketing actions’ results by providing a deeper appreciation of the heterogeneity within portfolios.

This white paper, therefore, delves into the benefits of understanding *uncertainty*, its applications, and practical considerations for end customers. All illustrations and results presented in this paper are derived from an internal Ethias dataset. We will also explore how the uncertainty measures discussed in this paper (Epistemic vs Aleatoric, Conformal) can be useful in managing the uncertainty of Large language models (LLMs) and their propensity to *hallucinate*.

*Akash.singh@uliege.be
†Ashwin.ittoo@uliege.be
‡Pierre.ars@ethias.be

1 Introduction

The world is currently experiencing the AI revolution, with advancements like self-driving cars (see *e.g.*Myers [May 2022]) and large language models (see *e.g.*OpenAI [2022]). This revolution is fundamentally transforming how data is collected and processed, driving unprecedented levels of efficiency, innovation, and transformation across all industries.

The financial sector, including banks and insurance companies, is actively embracing the opportunities presented by this remarkable development, particularly the power of machine learning tools. For instance, these tools enable insurance companies to:

1. **Set fairer premiums**: AI enables optimal premium calculations based on the most accurate risk and cost assessments related to an insurance contract. This ensures that customers benefit from premiums that accurately reflect their risk levels.
2. **Optimize customer engagement**: Customer engagement is a key factor in ensuring customer satisfaction. It manifests in various ways such as determining the most appropriate Next Best Action (NBA) to reward loyal customers, retain those with a high likelihood of churning, encourage cross-selling, ...

Traditional machine learning (ML) models, which do not account for *uncertainty*, primarily generate *point-wise predictions*, offering a single, definitive outcome. While these models have demonstrated remarkable accuracy across various applications and domains, they often fall short in providing critical context, namely the degree of (un)certainty with their predictions. A model that incorporates *uncertainty* provides not only the prediction but also information about the *distribution of the prediction*.

ML models are, of course, trained to achieve the most accurate predictions. However, overlooking the uncertainty inherent in these predictions can significantly impact decision-making, particularly in risk-sensitive industries such as insurance. Therefore, it is crucial to account for uncertainty before undertaking any action aimed at improving customer service. In this context, *uncertainty* can become a valuable asset for enhancing the quality of customer service.

Moreover, incorporating uncertainty into the modeling process has important implications at the management level :

1. It allows for a more finer interpretation of results.
2. By understanding the predictions and their associated uncertainty, as well as their impact on potential actions, insurance companies can gain a deeper insight into the risks associated with their predictions and develop robust strategies to mitigate those risks. This Uncertainty Quantification (UQ) empowers them to make more informed decisions regarding liability assessment, risk management, and resource allocation.

New regulations such as the EU AI Act and UK CFA Act emphasize on trustworthy models, which make fair and unbiased predictions. The EU AI Act, for instance, implements this requirement by classifying AI systems into four risk categories: unacceptable, high, limited, and minimal. Each category has specific regulations and compliance requirements for organizations developing or using those systems. One approach to ensure models' trustworthiness and fairness is through uncertainty quantification (UQ). Predictions made without any formalization or quantification of the associated uncertainty are not considered as trustworthy (Abdar et al. [2021], Hoffmann et al. [2021] and Kline [1985] .)

As ML models become increasingly sophisticated, understanding and managing uncertainty is becoming crucial for their reliable and responsible deployment across various domains(section 4). One notable example is Large Language Models (LLMs *e.g.* ChatGPT). UQ can help evaluate the reliability of outputs (words/sentences) predicted by these models. In doing so, it provides insights into the issue of hallucination (see section 4), which is essential for making these models more trustworthy and safer for deployment.

In this paper, we explore how appropriate modeling and handling of uncertainty can help insurance companies provide more efficient customer service through improved marketing management functionalities. As a use case, we consider identifying customers at risk of churning.

We will present two different approaches to quantify *uncertainty* around predictions:

1. **Uncertainty quantification, HEUQ** We introduce a method called Heterogeneous Ensemble for Uncertainty Quantification (HEUQ[1]) that leverages an ensemble of machine learning models (e.g., Logistic Regression, Neural Networks, Gradient Boosting Machines, Bagging, Random Forests, and CatBoost) to estimate uncertainty (Section 2). HEUQ essentially assesses the disagreement among these models to quantify the reliability of the predictions.

2. **Calibrating predictions, Conformal predictions** We will also explore a powerful technique called *conformal prediction*[2] which transforms raw predictions into statistically sound ones, guaranteeing a pre-defined level of confidence. (Section 3)

In addition to those applications, given the recent advent of LLMs, we provide an overview of UQ in LLMs, in Section 4.

2 Uncertainty

[1] see Singh et al. [July 2024]

[2] see Tibshirani [Spring 2023]

2.1 Beyond Machine Learning Prediction: Why Uncertainty Matters in Machine Learning

2.1.1 What is uncertainty?

Even the most sophisticated models can't be 100% accurate, and there will always be some degree of inherent uncertainty in their predictions. This uncertainty is often subdivided into (Quoted definitions are from Hüllermeier and Waegeman [2021]):

- **Data or Aleatoric uncertainty** "Roughly speaking, aleatoric (aka statistical) uncertainty refers to the notion of randomness, that is, the variability in the outcome of an experiment which is due to inherently random effects. The prototypical example of aleatoric uncertainty is coin flipping: The data-generating process in this type of experiment has a stochastic component that cannot be reduced by any additional source of information (except Laplace's demon). Consequently, even the best model of this process will only be able to provide probabilities for the two possible outcomes, heads and tails, but no definite answer".

- **Model or Epistemic uncertainty** "Uncertainty caused by a lack of knowledge (about the best model). In other words, it refers to the ignorance of the agent or decision maker, and hence to the epistemic state of the agent instead of any underlying random phenomenon. As opposed to uncertainty caused by randomness, uncertainty caused by ignorance can in principle be reduced based on additional information. For example, what does the word "kichwa" mean in the Swahili language, head or tail? The possible answers are the same as in coin flipping, and one might be equally uncertain about which one is correct. Yet, the nature of uncertainty is different, as one could easily get rid of it."

"In other words, epistemic uncertainty refers to the reducible part of the (total) uncertainty, whereas aleatoric uncertainty refers to the irreducible part."

Those are theoretical definitions, but let's note there is no clear-cut distinction between aleatoric and epistemic uncertainties, which are not really mutually exclusive (see *e.g.* Fox and ülkümen. [2011]).

Thus Total uncertainty is the gross sum of Aleatoric and Epistemic uncertainty.

Why should uncertainty matter?

- **Misinformed actions** If we blindly trust a prediction without considering its uncertainty, we might take actions that are ineffective or even harmful. For example, a misdiagnosis based on a flawed model for classifying cancer could lead to serious consequences.

- **Wasted resources** Ignoring uncertainty can result in wasted resources if we invest heavily in marketing campaigns or actions targeting customers with unreliable predictions.

By understanding the level of uncertainty associated with a prediction, we can better weigh the risks and benefits of the actions and decisions based on those predictions, ultimately benefiting both the customers and the insurance company.

To illustrate this, consider an insurance company aiming to control the churn rate (*i.e.*the percentage of customers leaving the company). Customers or policyholders are represented by a profile **x** where **x** are known features of customers like age, postal code, product information, etc. A key component is using a machine learning model to predict the likelihood of churn. The machine learning model learns a hypothesis (or model) h that estimates the probability of churn ($P(C|h,\mathbf{x})$). We can define the *uncertainty* as the level of confidence the company has in the prediction $P(C|h,\mathbf{x})$. A high probability of churn of $P(C|h,\mathbf{x}) = 80\%$ can be associated with either high confidence (low uncertainty) or low confidence (high uncertainty), depending on the performance of the model h for the profile **x**. High uncertainty could arise from limited data for similar profiles to **x**, leading to uncertain predictions for those profiles.

For personalized customer service, it is essential to consider an individualized assessment of uncertainty for each data point i.e. customer profile (**x**). Understanding this uncertainty is critical for the company. It allows for more targeted management of customers at risk of churning (e.g., prioritizing those with a high probability of churning and low uncertainty) or customers who are at minimal risk of churning (e.g., offering loyalty programs). This improves the accuracy of churn predictions and retention rates while enhancing personalized customer service.

2.2 Uncertainty quantification

Our study adopts the categorization by Hüllermeier and Waegeman [2021] of total uncertainty in machine learning into two key categories: aleatoric (or data) uncertainty and epistemic (or model) uncertainty:

$$totaluncertainty = aleatoricuncertainty + epistemicuncertainty.$$

Aleatoric uncertainty arises from inherent variability or noise within the data itself.

Epistemic uncertainty reflects limitations in the model's ability to capture the true underlying relationships in the data.

Several factors contribute to total uncertainty in predictions such as noise, variability, limited data, model complexity, and choice of models.

The distinction between these two key categories can then be made by considering an ensemble of N models $h_1, h_2, ..., h_N$, each providing a different prediction[3] $P(C|h_i,\mathbf{x}), 1 \leq i \leq N$.

This diversity allows the ensemble to capture a wider range of the inherent variability and complex relationships within the data. Consequently, we theorize that by

[3] $P(C|h_i,\mathbf{x})$ denotes the probability of churn as predicted by the learner h_i (one among the N : RF, LR,...) for the profile/customer with characteristics **x**.

learning a collection of diverse hypotheses, heterogeneous ensembles will have more confidence in their predictions or, at the very least, provide efficient ways to estimate aleatoric and epistemic uncertainties.

The figure 1 represents an heterogeneous model with three different learners.

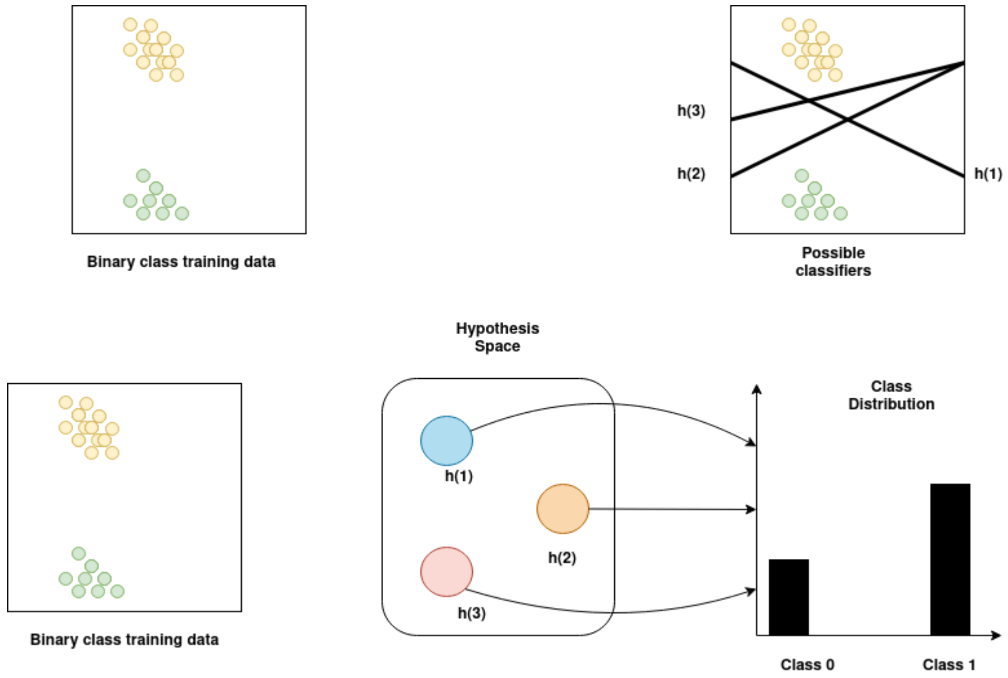

Figure 1: An example of the binary classification task. Three different optimal hypotheses are represented as three lines that divide the data into two classifying regions. The optimal hypothesis is the one that minimizes the overall error. The illustration also represents how multiple optimal hypotheses can also output similar class distributions.

The idea is therefore to use sufficiently heterogeneous machine learning models which due to their varied optimization functions will learn diverse "relationships" in the data (linear, nonlinear, based on trees or neural network,...). This diversity allows multiple ways of classifying a given profile. This multiplicity can then be used to dissect the "uncertainty" around the given profile into its two components: aleatoric and epistemic uncertainty.

For example, our (ensemble) HEUQ model, in its most comprehensive version, is based on six learners: h_1 = LR, h_2 = RF,h_3 = Bagging, h_4 = GBM, h_5 = catboost and, h_6 = NN.

This HEUQ model leads to the estimation of event probability computed according to the following formula :

$$P(C|HEUQ,\mathbf{x}) = \frac{P(C|h_1,\mathbf{x}) + P(C|h_2,\mathbf{x}) + P(C|h_3,\mathbf{x}) + P(C|h_4,\mathbf{x}) + P(C|h_5,\mathbf{x}) + P(C|h_6,\mathbf{x})}{6}$$

Currently, we quantify the disagreement between the learners(models) through the

Epistemic Uncertainty denoted by $u_e(\mathbf{x})$ by measuring divergence using the KL-Divergence method [4].

As discussed previously *Epistemic Uncertainty* is a component of *Total Uncertainty* (denoted by $u_t(\mathbf{x})$).

We can view $u_t(\mathbf{x})$ as the loss function value obtained by the algorithm during optimization (the process that determines the model's parameters). Traditionally, however, total uncertainty has been defined through the entropy of the predictive probability[5]. Finally, the *Aleatoric Uncertainty*, denoted by $u_a(\mathbf{x})$, is calculated as the difference between $u_t(\mathbf{x})$ and $u_e(\mathbf{x})$.

We then have the fundamental formula :

$$u_t(\mathbf{x}) = u_a(\mathbf{x}) + u_e(\mathbf{x}). \tag{2}$$

An example of the behavior of $u_a(\mathbf{x})$ vs $u_e(\mathbf{x})$ is given by the figure 2 based on HEUQ results on a internal churn dataset.

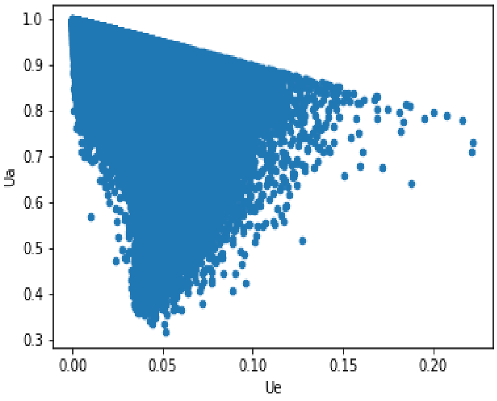

Figure 2: Epistemic Uncertainty(Ue,X-axis) vs Aleatoric Uncertainty(Ua,Y-axis)

2.3 Actionable insights: Uncertainty analysis and the marketing management!

The information provided by the HEUQ-estimations of $u_e(\mathbf{x})$, $u_a(\mathbf{x})$ and $u_t(\mathbf{x})$ (the sum of the two previous ones) can be translated into valuable *Actionable insights* for marketing management.

The approach is to consider individual profiles $\mathbf{x}$ and examine the behavior of total uncertainty, $u_t(\mathbf{x})$, epistemic uncertainty $u_e(\mathbf{x})$ and aleatoric uncertainty $u_a(\mathbf{x})$.

[4]

$$u_e(\boldsymbol{x}) = \frac{1}{6}\sum_{j=1}^{6} KL[p(\cdot|h_j, \boldsymbol{x}), p(\cdot|HEUQ, \boldsymbol{x})] \tag{1}$$

Where $KL[q_1, q_2]$ denotes the KullBack-Leibler divergence between (here) discrete binary probability distributions q_1 and q_2.

[5] For an ensemble method, like $HEUQ$, and a profile/customer $\mathbf{x}$, the Total uncertainty is here defined by the entropy $H(P(C|HEUQ, \mathbf{x})) = P(C|HEUQ, \mathbf{x}) \; log_2(P(C|HEUQ, \mathbf{x}) + P(notC|HEUQ, \mathbf{x}) \; log_2(P(notC|HEUQ, \mathbf{x})$, where $notC$ denotes the event *"the customer does not churn"*.

2.3.1 Optimised sub-portfolio: Precision metric optimization

A key aspect of marketing management is predicting, with reasonable precision, the probability that a customer **x** will churn.

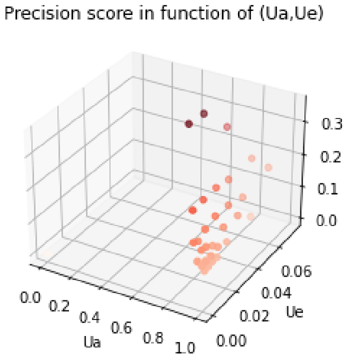

Figure 3: Precision score, in the function of classes defined from the crossing of quantile bins (ordered by increasing level of uncertainty), for large enough classes, representing at least 0.5% of the data. (for statistical reasons)

Given the available features **x** and domain-specific information/parameters available, the marketing manager can assign a "profile value" $V(\mathbf{x})$ and decide on the appropriate strategy (e.g. offering an additional discount, providing extra coverage at a lower cost, etc...).

To evaluate the cost and benefit of the strategy, the marketing manager will first select a sub-portfolio for which the probability of churn is well established. Specifically, he/she will choose a sub-portfolio that maximizes the ***precision*** *score* defined by

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

.

In churn prediction, the precision score indicates the proportion of customers predicted to churn (positive cases) who actually do churn (true positives or TP) for the sub-portfolio. The false positives (FP) represent customers predicted to churn who, in reality, remain with the company.

Now, let's explore the relationship between precision scores, uncertainties, and decision making.

1. HEUQ should achieve a lower precision score for high values of $u_t(\mathbf{x})$ or $u_a(\mathbf{x})$.
2. Very high epistemic uncertainty suggests significant disagreement among the different hypotheses (models/learners within HEUQ).
3. In such cases, management should exercise caution when making any (marketing) decision based on the predictions.

4. Conversely, very low epistemic uncertaintymay also require caution, as the agreement among models could stem from low data density and may not indicate high confidence in the predictions.

For these reasons, when we fix an interval of $u_a(\mathbf{x})$ values, we often see a sort of concave pattern of the precision score in the function of $u_e(\mathbf{x})$: precision initially increases for lower uncertainty values, reaches a peak, and then decreases (see figure 3). This pattern, when confirmed on a specific dataset, can be valuable for marketing management in determining "optimal sub-portfolios".

2.3.2 Optimized sub-portfolio: Balanced accuracy optimization

Balanced accuracy (BA) is a metric used to assess the performance of a classification model and is defined as follows :

$$BA = \frac{TPR + TNR}{2},$$

where $TPR = \frac{TP}{TP+FN}$ and $TNR = \frac{TN}{FP+TN}$.

We now demonstrate how the HEUQ uncertainty decomposition can be utilized to build sub-portfolios with a high BA.

Ue increasing →				**Balanced accuracy**				
				80.01	78.58	82.51	84.64	79
	63.89	74.35	69.24	69.72	70.84	73.24	71.55	70.3
Ut increasing ↓	60.25	68.08	64.41	67.73	67.41	65.45	65.39	65.72
	58.42	63	62.64	58.92	62.54	63.2	64.59	58.38
	55.27	57.26	57	57.6	55.92	60.71	53.72	56.78
	53.15	52.66	50.45	53.32	52.5	53.22	49.18	52.31

Table 1: Balanced accuracy (BA) for sub-portfolios obtained by crossing of the 8-quantiles for Ue and 6-quantiles for Ut. For example, 62.54 is the BA for the sub-portfolio composed by profiles satisfying $u_e(\mathbf{x})$ being between $4th$ 8-quantile and $5th$ 8-quantile and $u_t(\mathbf{x})$ being between $3th$ 5-quantile and $4th$ 5-quantile. The marketing manager can then select subportfolios based on $u_e(\mathbf{x})$ an $u_t(\mathbf{x})$ according to his objectives in terms of BA.

Intuitively, high epistemic uncertainty $u_e(\mathbf{x})$ indicates disagreement among the different models/learners. We can therefore hypothesize that HEUQ, as an average of these learners, will yield better BA in sub-portfolios selected based on relatively high levels of $u_e(\mathbf{x})$ (table 2) combined with relatively low levels of $u_t(\mathbf{x})$. However, this observation is dataset-specific and should be interpreted with caution.

This type of *intricate* behavior can be can be better understood by examining table 2 which was constructed as follows :

1. For any of the uncertainties U

2. for a set of quantiles (q) Q
3. we select the sub-portfolio S(U,q) with $U(x) \geq quantile(U, q)$
4. we compute BA for S(U,q)

Example ((see table 2) :

1. $U = u_e$
2. $q = 20\%$: that means that we take the profiles **x** having an $u_e(\mathbf{x})$ value among the 80% highest $u_e(\mathbf{x})$ values of the total portfolio, that is greater than the value given by *Unc Minimal Bound*, here 0.0062
3. For this sub-portfolio, the observed BA is 66.96

It's important to note that this behavior is not the same for other sources of uncertainty, such as total uncertainty $u_t(\mathbf{x})$ and aleatoric uncertainty $u_a(\mathbf{x})$.

Our methodology enables more refined strategies. For instance, by combining $u_e(\mathbf{x})$ with $u_t(\mathbf{x})$, we can optimize a portfolio (w.r.t to BA for example) by considering the behavior of BA observed as $u_e(\mathbf{x})$ values increase within a given range of $u_t(\mathbf{x})$ values. To illustrate this point, let's look at table 1, specifically the first row, which corresponds to the lowest values of $u_t(\mathbf{x})$):

1. We observe an approximately concave behavior , where BA initially increases and then decreases.
2. This is followed by a sharp decline in BA for the highest values of $u_e(\mathbf{x})$.

This set of strategies, only sketched here, can be represented in a scheme, see figure 4.

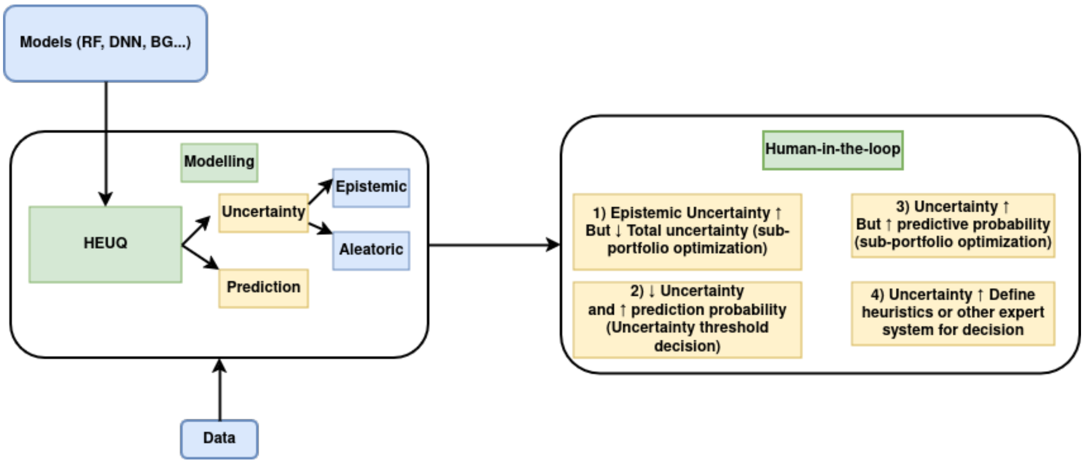
Models (RF, DNN, BG...)
Modelling
HEUQ
Uncertainty
Epistemic
Aleatoric
Prediction
Data
Human-in-the-loop
1) Epistemic Uncertainty ↑ But ↓ Total uncertainty (sub-portfolio optimization)
2) ↓ Uncertainty and ↑ prediction probability (Uncertainty threshold decision)
3) Uncertainty ↑ But ↑ predictive probability (sub-portfolio optimization)
4) Uncertainty ↑ Define heuristics or other expert system for decision

Figure 4: HEUQ Framework Human-in-the-loop.

Uncertainty type	Datasize in %	Quantile	Unc Minimal Bound	BA
u_e	100%	0.00	0.0000	**64.95**
u_e	90%	0.10	0.0035	**66.01**
u_e	80%	0.20	0.0062	**66.96**
u_e	70%	0.30	0.0095	**68.07**
u_e	60%	0.40	0.0139	**69.67**
u_e	55%	0.45	0.0167	**70.34**
u_e	50%	0.50	0.0201	**71.26**
u_e	45%	0.55	0.0242	**71.98**
u_e	40%	0.60	0.0292	**72.96**
u_e	30%	0.70	0.0405	**74.34**
u_t	100%	0.00	0.3678	64.95
u_t	90%	0.10	0.7246	62.81
u_t	80%	0.20	0.8114	61.04
u_t	70%	0.30	0.8692	59.59
u_t	60%	0.40	0.9132	58.40
u_t	55%	0.45	0.9300	57.63
u_t	50%	0.50	0.9444	56.84
u_t	45%	0.55	0.9563	56.11
u_t	40%	0.60	0.9662	55.66
u_t	30%	0.70	0.9818	53.98
u_a	100%	0.00	0.3164	64.95
u_a	90%	0.10	0.6690	62.85
u_a	80%	0.20	0.7603	61.15
u_a	70%	0.30	0.8283	59.72
u_a	60%	0.40	0.8829	58.47
u_a	55%	0.45	0.9047	57.84
u_a	50%	0.50	0.9225	57.03
u_a	45%	0.55	0.9373	56.47
u_a	40%	0.60	0.9493	55.76
u_a	30%	0.70	0.9677	54.44

Table 2: Evolution of the BA score with increasing level of uncertainty $(u_e(\mathbf{x}), u_t(\mathbf{x}), u_a(\mathbf{x}))$.

Of course, the average distance (AD) between the learner's estimation of event probability (*e.g.* here churn) and the threshold is a crucial factor for explaining the BA levels. Our statistical results indicates that both $u_e(\mathbf{x})$ and $u_a(\mathbf{x})$ improve data fitting (*e.g.* the AIC in a GLM test): this improvement in fitting further supports the use of uncertainty parameters in management decisions.

These observations are specific to the specific Ethias portfolio considered in this study and may not be directly applicable to other portfolios.

2.3.3 HEUQ for human-in-the-loop decision making

Epistemic uncertainty (u_e) in HEUQ reflects the disagreement among different base learners. HEUQ's uncertainty estimates can serve as a valuable guide for human analysts, helping them make informed decisions about NBA (Next Best Action) for profiles/customers with high $u_e(\mathbf{x})$. For such profiles, where the model's confidence is lower, automated solutions may be less reliable.

3 Conformal prediction

In the preceding section, we discussed how to implement an uncertainty decomposition to help marketing managers develop effective strategies based on a thorough understanding of prediction uncertainty. However, perfect predictive models do not exist, and even the best models cannot eliminate all uncertainty surrounding a prediction.

The Conformal Prediction (CP) bridges this gap by providing calibrated uncertainty sets. [6] Unlike point estimates, CP guarantees that the true value lies within the predicted set with a selected confidence level (e.g., 90%).

3.1 Advantages of Conformal predictions

- Calibrated Uncertainty: CP quantifies the reliability of predictions, enabling more informed decision-making.
- Robustness to Outliers: CP is less affected by outliers compared to traditional methods, leading to more reliable results.
- Improved Risk Management: By understanding the prediction uncertainty, users can better assess potential risks associated with model outputs.

3.2 Applications of Conformal Prediction

- Safety-Critical Systems: CP is well-suited for scenarios where reliable predictions are essential, such as autonomous vehicles or medical diagnosis.
- Active Learning: CP can guide active learning algorithms by prioritizing data points that will most significantly reduce prediction uncertainty.
- Cost-Sensitive Applications: CP helps optimize resource allocation by focusing effort on high-confidence predictions with the greatest potential impact.

3.3 Implementing Conformal Prediction

This section explores specific techniques for implementing CP with various machine learning models (e.g., Classification). Key considerations include:

- Choice of Calibration Set: Strategies for selecting a calibration set to ensure reliable uncertainty estimation.
- Computational Efficiency: Address potential computational costs associated with CP algorithms.

[6] A good technical introduction to CP is provided by Tibshirani [Spring 2023].

- Interpretability of Conformal Prediction sets: Explore methods for visualizing and interpreting the predicted uncertainty sets to enhance user understanding.

3.4 Conformal prediction in the binary classification case

A classification model assigns a profile to one of several discrete classes. For instance, in a binary churn model, the HEUQ algorithm classifies a customer **x** into either a *churn class* (likely to churn) or *not churn class* (likely to not churn).

Generalizing the classical concept of hypothesis testing, conformal prediction aims to determine the p-value for each possible class: if this p-value exceeds a threshold (related to the 'significance level (α)' or 'target miscoverage level'), the label is included in the prediction set.

The threshold is computed through the $\alpha - quantile$ of a set of likelihood scores.

For a binary $0/1$ classification problem, the "Conformal Prediction set", can be one of the following four possibilities: $\emptyset, \{O\}, \{1\}, \{O, 1\}$.

To illustrate these concepts, consider concrete examples of "Conformal Prediction sets", in the binary classification context of Ethias churn dataset (The event churn, denoted by C, corresponding to the $'1'$ class), with the following parameters and notations :

1. Dataset: Ethias churn dataset
2. Significance level : $\alpha = 0.1$
3. corresponding threshold: 0.3
4. let's $P(C|\mathbf{x})$ denote the churn probability provided by the predictive model (e.g.HEUQ or any other learner like Catboost) for the profile **x**.
5. $CC(\mathbf{x})$ is the "Conformal Prediction set (or class)" for the profile **x**.

The crucial result is that, under mild conditions (mainly exchangeability), it can be proven that the final decision regarding the profile **x** (churn or not churn) belongs to the Conformal Prediction set $CC(\mathbf{x})$ with a probability of at least $1 - \alpha$. This result applies globally to the entire portfolio for both classes.

To illustrate how CP works, consider three different profiles $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$, each falling in a different "Conformal Prediction set" :

- $P(C|\mathbf{x}_1) = 0.15$

   (a) Then $P(C|\mathbf{x}_1) \leq 0.30$ and $1 - P(C|\mathbf{x}_1) \geq 0.30$

(b) Therefore $CC(\mathbf{x}_1) = \{O\}$, and the marketing manager can be confident at the level $\alpha = 0.1$ with the the (conformal) prediction that $\mathbf{x}_1$ will effectively not churn, that is :

$$P(\mathbf{x}_1 \ \ will \ \ not \ \ churn) \geq 0.90$$

- $P(C|\mathbf{x}_2) = 0.75$

  (a) Then $P(C|\mathbf{x}_2) \geq 0.30$ and $1 - P(C|\mathbf{x}_2) \leq 0.30$

  (b) Therefore $CC(\mathbf{x}_2) = \{1\}$, and the marketing manager can be confident at the level $\alpha = 0.1$ with the (conformal) prediction that x_2 will churn, that is :

$$P(\mathbf{x}_2 \ \ will \ \ churn) \geq 0.90$$

- $P(C|\mathbf{x}_3) = 0.60$

  (a) Then $P(C|\mathbf{x}_3) \geq 0.30$ and $1 - P(C|\mathbf{x}_3) \geq 0.30$

  (b) Therefore $CC(\mathbf{x}_3) = \{O, 1\}$, so we do not have enough evidence to definitively classify $\mathbf{x}$ in the subsets of either churners or non-churners with a confidence at least equal to 0.90.

The Conformal Prediction set $\emptyset$ can occur when, for example, the threshold is 0.60 and the prediction probability $P(C|\mathbf{x}) = 0.55$.

The figure 5 represents an illustration of the Conformal Prediction Process.

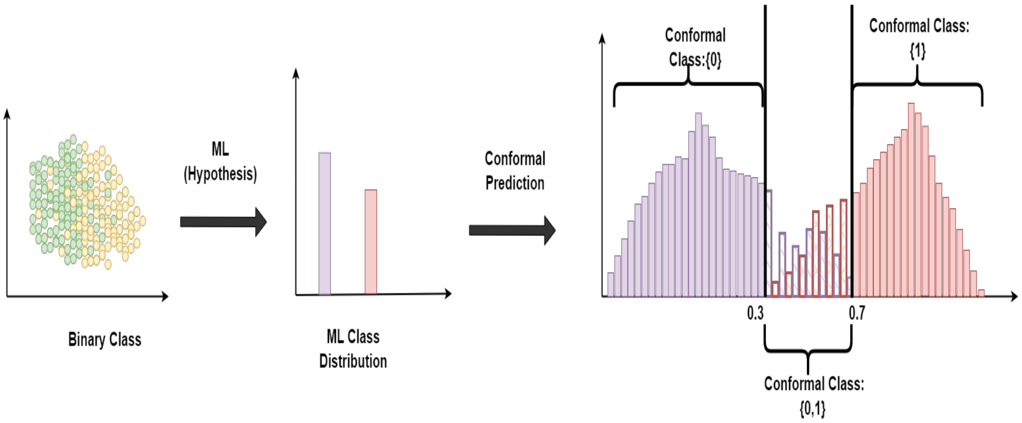

Figure 5: An example of binary conformal prediction approach with a threshold of 0.30. On the right figure, in abscissa, we have the predicted probability of churn ($P(C|x)$). With such a threshold, there is no possibility for the selection $\emptyset$ as a Conformal Class. Here ”ML” denotes any machine learning algorithm: random forest, neural network, catboost, HEUQ,...

Of course, an error in classifying an insured as a *non-churner* has different implications for both the insurance company and the insured compared to the opposite error.

Therefore Vovk [2013] extended this result to a *Class Conditional Validity*, ensuring that the confidence level is respected for both classes: *non churners* and *churners.*

3.5 How can Conformal Prediction be useful in improving marketing management ?

The Conformal Prediction approach allows decision-makers to take decisions based on reduced uncertainty. For example, consider a fraud portfolio manager. If, for instance, the Conformal Prediction set (henceforth referred to as Conformal-class) is $\{O\}$ (resp. $\{1\}$) and the significance level is 0.1, then the manager can make decisions about the case with confidence, knowing that the fraud probability is less than 10%. (or at least 90%.) Cases classified as $\{O, 1\}$require further analysis, potentially by a *human expert*.

The same is true for a marketing manager in charge of the churn portfolio, but for the difficulty, in real life, to transfer the decision to a *human expert*. Nevertheless, the process can be useful for efficiently allocating the marketing budget to cases with significant uncertainty reduction, thereby avoiding wasted time, money, or customer frustration from being contacted ”by mistake.”

To illustrate this, we apply **Conditional version** of the Conformal Predictor approach to the Ethias churn portfolio.

Due to confidentiality, we cannot fully disclose the results comparing ’churners’ and ’non-churners.’ However, to illustrate the relevance of CP, we present the following noteworthy results for a *Confidence level* of 4%:

1. The CP provides a conformal class of $'1'$ with an observed churn ratio $\geq 25\%$ in the testing set. The Marketer can then take specific actions for these *potential churners* with a relatively high level of confidence.
2. The CP provides a conformal class of $'0'$ with an observed churn ratio $\leq 2.5\%$ in the testing set. The marketer can offer these customers specific loyalty programs or cross-selling opportunities.
3. In all cases, these customers will receive attractive proposals, potentially leading to an improved cover-to-premium ratio.
4. The results can be monitored by the Marketer and aligned with management goals by choosing an appropriate *Confidence level*, which will affect the size and composition of the different Conformal Prediction Classes.

4 Declinations of uncertainty in the context of Large Language Models

Large language models (LLMs) essentially model a probability distribution over text by combining language elements (or ”sub-claims”) through complex composition.

Depending of the application or context, the accuracy and reliability of LLM outputs can be critical.

Therefore, managing the uncertainty surrounding an LLM's outputs is essential. In this section, we aim to show that the tools discussed in the previous sections can be applied to analyzing the reliability of language generation.

4.1 Epistemic and Aleatoric concepts in the context of LLMs

Here, we explore how the concepts of epistemic and aleatoric uncertainties can help improve the reliability of outputs generated by LLMs.

While LLMs are highly efficient and show great potential, it is well known that they are prone to hallucinations, which presents a significant challenge. Certain ”uncertainty” measures can be derived from next-token prediction probabilities, which have been shown to be well-calibrated in multiple-choice question-answering settings (see, for example, Harvard [April 2024]).

In Harvard [April 2024], the authors highlighted the relevance of distinguishing between *aleatoric* and *epistemic* uncertainties. They proposed a promising approach, stating that:

1. “Supervised linear probes trained on a language model's internal activations can achieve high accuracy at classifying epistemic versus aleatoric uncertainties, even when the probes are evaluated on unseen text domains (probes trained on Wikipedia text, for example, generalize well to code data).”
2. “One way to define the difference between epistemic and aleatoric uncertainty is that aleatoric uncertainty is inherent to the randomness in language, while epistemic uncertainty is “in the eyes of the beholder.” As the quantity of training data and computation time increases, models will learn more of what is knowable, and epistemic uncertainty will recede. In other words, as language models become bigger and trained for longer, they become better proxies for the “true” distribution. ”

But how can we extend the notions of Total, Aleatoric, and Epistemic uncertainties to the context of Large Language Models (LLMs)?

Although challenging, it is possible:

1. **Aleatoric Uncertainty**: Measuring aleatoric uncertainty in LLMs is more complex than in traditional classification tasks. In classification, uncertainty arises from the choice between distinct alternatives, which can often be explained by available features. However, in LLMs, the input text may be ambiguous, incomplete, or open to multiple interpretations. This type of uncertainty is inherently aleatoric because it cannot be reduced or eliminated, even with additional data or improved models.
2. **Epistemic Uncertainty**: Identifying and quantifying epistemic uncertainty in LLMs presents significant challenges, especially when the model encounters

out-of-distribution or unfamiliar data. Epistemic uncertainty reflects the model's lack of knowledge and can be reduced with better training or more data. However, due to the complexity of language tasks, detecting this uncertainty in LLMs can be difficult. .

3. **Defining Uncertainties**: Despite these challenges, some approaches have been developed to define Total, Aleatoric, and Epistemic uncertainties in the context of LLMs. Building on a predictive model ppp, these definitions extend classical concepts (see, for example, Hou et al. [July 2024]):
   (a) **Total uncertainty:** This is represented by $H(p)$ the entropy[7] of p.
   (b) **Aleatoric uncertainty:** This is the entropy of the ground-truth probablity q, representing the irreducible uncertainty due to the inherent randomness in the data.
   (c) **Epistemic uncertainty:** This is calculated as the difference between Total and Aleatoric uncertainties. It reflects the uncertainty in the model itself, which could potentially be reduced with more data or better training.

   An output with a large systemic entropy is then judged unreliable and rejected.

4.2 Conformal prediction in the context of LLMs

Conformal prediction (CP) can also offer effective ways to mitigate issues such as *hallucinations* in large language models (LLMs).

Consider a standard language model L that generates an output $L(X)$ from an input X.

To ensure performance guarantees, some extensions of CP have been developed.

To provide performance guarantees, some extensions of CP have been developed. We present two approaches:

1. Using an approach similar to CP, Quach et al. [June 2024] "calibrated a stopping rule for sampling LM outputs that get added to a growing set of candidates until they are confident that the set covers at least one acceptable response. Since some samples may be low-quality, they also simultaneously calibrate a rejection rule for removing candidates from the output set to reduce noise."

2. Christopher Mohri [February 2024] adapted the CP algorithm by defining an *Entailment Operator* $E(\cdot)$ assigning to each X a subset of claims entailing $L(X)$. A concept similar to the CP prediction set can then be defined by $E(L(X))$, extending the traditional covering properties of CP to this setting, and allowing the rejection of claims that are insufficiently reliable.

[7] For a discrete random X taking the values $(x_1, \cdots, x_n)$ with the probabilities $(p(x_1), \cdots, p(x_n))$, the entropy $H(X)$ is defined by $H(X) = -\sum_{i=1}^{n} \; p(x_i) \; log_2(p(x_i))$.

However, these methods have limitations (see John J. Cherian [June 2024]) :

1. the true probablity of correctness, even when using to the CP framework, can vary significantly depending on the characteristics of the input X.
2. Additionally, the number of rejected claims can be excessively high, rendering these methods practically inefficient.

To address the *impracticality* of these approaches, Candes et al. (2024) introduced two methods:

1. **Conditional boosting :** By differentiating through the conditional conformal algorithm, this method increases the retention of claims while maintaining performance.
2. **Level-adaptative conformal prediction :** This is a well *calibrated*[8] approach that allows the validity of the conformal output to depend on the characteristics of the input X.

5 Conclusion

In this white paper, we have explored the critical role of uncertainty quantification (UQ) in enhancing the reliability and personalization of customer service, particularly within the insurance industry. By integrating machine learning models, such as the Heterogeneous Ensemble for Uncertainty Quantification (HEUQ) and Conformal Prediction (CP), we demonstrated how uncertainty can be systematically quantified and utilized to drive informed decision-making.

5.1 Key Takeaways

- **Enhanced Decision-Making** By understanding both epistemic (model-based) and aleatoric (data-based) uncertainties, insurance companies can make more informed decisions, improving customer engagement strategies, such as Next Best Action (NBA).
- **Improved Customer Service** Incorporating uncertainty measures, including conformal predictions, allows for a more personalized approach to customer service. This not only helps in identifying customers at risk of churn, as with any predictive model, but also in evaluating the uncertainty surrounding the prediction. This leads to more targeted interventions and higher retention rates.

[8] As precised by John J. Cherian [June 2024] , "Calibration requires that the true probability of correctness matches the issued one. For example, if a weather forecaster claims that there is a 70% chance of rain, their forecast is calibrated if it actually rains for 70% of the days on which a 70% forecast is issued.

- **Risk Management** The methods discussed offer robust tools for risk management, allowing companies to better understand the potential risks associated with their predictions and to develop strategies to mitigate these risks effectively.

As ML tackles increasingly complex problems and is widely adopted, robust uncertainty quantification (UQ) becomes essential for building trust in model-driven decisions.

Organizations should focus on adapting UQ techniques to ensuring the reliability and trustworthiness of ML-powered choices.

By systematically incorporating advanced UQ methodologies, organizations can not only refine their decision-making processes but also enhance the overall customer experience.

We also highlighted how the uncertainty approaches presented in this paper (Aleatoric vs Epistemic uncertainties; conformal prediction) can be valuable in improving the reliability of outputs from large language models (LLMs), thereby reducing the risk of *hallucinations*.

6 Contributors

Ethias

Tom To Hoang, Head of Data Office

François Dehouck, Head of Data Analytics

Pierre Ars, Actuarial Innovation Expert

HEC

Ashwin Ittoo, Full Professor ULiège— Co-Founder AIASHI

Akash Singh, PHD student, ULiège

NRB

Nicolas Dumazy, Chief Strategy and Data Officer, NRB Group

François Collienne, Head of Guild

Norman Marlier, Data Science Analyst

References

Andrew Myers, May 2022. URL https://hai.stanford.edu/news/how-ai-making-autonomous-vehicles-safer.

OpenAI. Chatgpt : Get answers. find inspiration. be more productive., 2022. URL https://openai.com/chatgpt/.

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021. ISSN 1566-2535. doi: 10.1016/j.inffus.2021.05.008.

Lara Hoffmann, Ines Fortmeier, and Clemens Elster. Uncertainty quantification by ensemble learning for computational optical form measurements. *Machine Learning: Science and Technology*, 2(3):035030, 2021.

S. J. Kline. The Purposes of Uncertainty Analysis. *Journal of Fluids Engineering*, 107 (2):153–160, 06 1985. ISSN 0098-2202. doi: 10.1115/1.3242449. URL https://doi.org/10.1115/1.3242449.

Akash Singh, Ashwin Ittoo, and Pierre Ars. Heterogeneous ensemble framework for uncertainty quantification (heuq) in operations research. *submitted for publication*, July 2024.

Ryan Tibshirani. Conformal Prediction , advanced topics in statistical learning, Spring 2023.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110: 457–506, 2021.

Craig R. Fox and Gülden ülkümen. Distinguishing two dimensions of uncertainty. In Brun, Keren, Kirkebøen, and Montgomery, editors, *Perspectives*on*Thinking,*Judging,*and*Decision*Making*. Oslo:Universitetsforlaget, 2011.

Vladimir Vovk. Conditional validity of inductive conformal predictors. *Machine Learning*, 92:349–376, 2013.

Kempner Institute Harvard, April 2024. URL https://kempnerinstitute.harvard.edu/research/deeper-learning/distinguishing-the-knowable-from-the-unknowable-with-language-models/.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhan. Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:2311.08718*, July 2024.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzil. Conformal language modeling. *arXiv preprint https://arxiv.org/pdf/2306.10193*, June 2024.

Tatsunori Hashimoto Christopher Mohri. Language models with conformal factuality guarantees. *arXiv preprint https://arxiv.org/pdf/2402.10978*, February 2024.

Emmanuel J. Candès John J. Cherian, Isaac Gibbs. Large language model validity via enhanced conformal prediction methods. *arXiv preprint https://arxiv.org/pdf/2406.09714*, June 2024.