Multi-channel multi-model customer conversion prediction in the insurance domain.*

Akash Singh[1][0000−0003−1679−8232], Ashwin Ittoo[1], Pierre Ars[2], and Elise Vandomme[1][0000−0001−6471−8968]

[1] HEC Liège - School of Management of the University of Liège, Liège, Belgium
Akash.Singh@uliege.be
https://www.hec.uliege.be
[2] Lead Actuarial Innovation, Ethias Insurance, Liège, Belgium

Abstract. This study addresses the challenge of predicting customer conversion across multiple sales channels within the insurance domain, an area that remains relatively underexplored compared to the extensive literature on e-commerce and online customer conversion. To bridge this gap, we evaluate the predictive capabilities of six popular machine learning (ML) models in estimating the likelihood of existing customers subscribing to additional services within 30 days of receiving an offer. The dataset utilized incorporates demographic profiles, vehicle specifications, portfolio compositions, and multi-channel engagement. The data exhibits varying degrees of class imbalance for customer conversion depending on the specific sales channel.

Our findings indicate distinct customer behaviors across sales channels, requiring channel-specific predictive models to optimize customer acquisition and conversion prediction. Among the evaluated models, CatBoost demonstrates superior performance, achieving a balanced accuracy (BA) of 82.20% with an F1 score of 87.7% on digital sales data and 73.66% BA with an F1 score of 73.6% for physical sales. While Neural Networks perform competitively (82.10% BA in digital and 72.70% BA in physical channels), CatBoost consistently outperforms by a marginal yet notable margin. The observed 10% performance disparity between channels reflects fundamental differences in customer behavior.

Keywords: Customer conversion prediction · Classification · Machine learning.

1 Introduction

Accurately predicting customer conversion is crucial for optimizing sales strategies, driving revenue, fostering growth, and enabling effective resource allocation, which makes it critical for business. The application of machine learning in predictive analytics has shown remarkable success in forecasting sales [9, 11, 12, 14]. While there has been a plethora of work, there remains no clear consensus on

* Supported by Ethias through the HEC Digital Labs.

the most effective model due to the wide variety of methodologies employed by researchers. Existing studies predominantly focus on e-commerce, online sales, or digital channels, leading to a research gap in the insurance industry [6, 11, 12, 14]. Furthermore, to the best of our knowledge, no study has examined customer conversion in the context of physical sales, and this may be attributable to difficulties in data collection. Additionally, there is an absence of comparative studies between different sales channels, i.e. physical and digital sales.

The objective of this research is to address the aforementioned gaps and to advance the current state of the art in Machine Learning (ML) applications, specifically in the context of customer conversion. Our focus is on predicting the likelihood that existing or new customers will subscribe to one or more insurance services when presented with an offer. We analyze historical customer data including user details, product subscriptions, demographic characteristics, and engagement channels to identify customers with a higher likelihood of subscription. This study demonstrates that channel-specific ML models can improve decision-making. In the context of insurance sales, this approach holds significant promise for understanding customer behavior across different sales channels and addressing the challenges posed by data imbalances in these channels.

The goal of this investigation is twofold. First, it aims to benchmark multiple ML models to identify the best performing one for predicting customer conversion sales. Second, it aims to assess conversion rates across physical and digital sales channels to estimate the reliability of channel-specific predictive models, while addressing the challenges of imbalanced data often encountered in different channels.

The study offers several important contributions. To the best of our knowledge, this study is the first to analyze customer conversion prediction across both physical and digital sales channels in the insurance domain, addressing a significant gap in existing research. Furthermore, our study focuses on the insurance domain, evaluating multiple machine learning models, identifying CatBoost and Neural Networks as the top performers while addressing class imbalance challenges.

2 Literature review

Yeo et al. [14] focus on e-commerce conversion prediction, introducing a joint modeling approach integrating customer and product-level conversion patterns based on the buying decision process. Their study employs XGBoost on a simulated dataset based on real-world data. Ji et al. [13] propose a time-aware recommendation model for customer conversion, integrating recommendation relevance and conversion time using survival analysis. Recommendation algorithms often fail to account for the mismatch between a product's actual conversion period and the target conversion period set by applications [13]. Tokuç et al. [12] leverage LightGBM for purchase intent prediction using e-commerce clickstream data, addressing high-dimensionality, class imbalance challenges, non-purchase sessions, and the temporal variability of user behavior. Martínez et al. [7] pro-

pose a predictive analytics framework for customer behavior in non-contractual settings, using gradient tree boosting to forecast future purchases. Their model, tested on a large dataset (10,000 customers and 200,000 transactions), achieves 89% accuracy and 0.95 Area Under the Curve (AUC)[3] for predicting next-month purchases. Balut et al. [9] introduce two models designed for real-time customer conversion prediction, aiding competitive cost-per-click bidding strategies in digital marketing campaigns, particularly for Google Ads. Soni et al. [10] evaluate the effectiveness of temporal, demographic, and behavioral factors in enhancing customer conversion. The authors employed the Decision Tree, Random Forest, XGBoost, and Gradient Boosting machine to evaluate factors influencing e-commerce customer conversion. Their study found that behavioral-based personalization most effectively enhances conversion rates. Bag et al. [1] examine how AI-driven digital engagement increases online sales and how satisfying online shopping experiences influence repurchase intentions. Using data from Indian consumers, their research indicates that AI positively affects user engagement and conversion, leading to satisfying user experiences.

Despite these advances, most studies focus on digital sales, with limited research on conversion prediction in physical sales settings. Additionally, comparative analyses between digital and physical sales channels remain scarce, highlighting the need for further research in this area.

3 Dataset and Preprocessing

The research paper leverages a historical customer dataset containing 190,000 records from a leading insurance provider in Belgium for car insurance. The dataset includes a rich set of features describing customer demographics, portfolio composition, vehicle specifications, and behavioral attributes. [4]

Customer demographics encompass data related to the characteristics of our customer base. 21.2% of customers are 25 years old or younger, 46.8% fall between 26 and 45 years old, 26.1% are aged between 46 and 65, and 5.9% are 66 years old or above. The gender distribution reveals a predominance of male customers (62.4%) compared to females (37.6%). Customers are also categorized into lifecycle segments such as young professionals and retirees, providing further granularity in demographic profiling.

Portfolio composition reflects the insurance products that customers are currently subscribed to, offering insights into their coverage preferences and product bundling behavior.

Vehicle specification includes specifications such as vehicle age and engine power. The average vehicle age is 7.5 years, with a standard deviation of 5 years; vehicle ages range from 0 to 60 years. Engine power distribution shows that 78%

[3] AUC indicates a model's ability to discriminate between different classes. It represents the probability that a randomly chosen positive instance will be ranked higher than a randomly chosen negative instance by the model [2].

[4] Due to the confidential nature of the data, we disclose only information that will not cause any prejudice to the insurer.

of registered vehicles have engines rated at 100 kW or less, 20.7% fall within the 101–200 kW range, and only 1.3% exceed 200 kW.

Additionally, the dataset incorporates behavioral variables such as whether customers have opted in to receive marketing communications, the number of insurance quotes requested in the past 30 days, and any history of insurance claims. These behavioral indicators provide important signals of customer intent and engagement.

We divide our data into two primary groups: Physical (Phy) and Digital (Web) channels. Physical selling refers to any form of customer engagement and offers that are made by a representative, such as face-to-face, or through a phone interactions. This can include insurance offers made directly in the insurance provider's office or over the phone with a representative assisting the customer. On the other hand, digital selling involves interactions that occur online, through digital platforms such as a company's website, mobile application, or through email marketing campaigns. These digital channels enable offers to be made without direct human interaction. In our data, we observed that the conversion rates in Physical channels is several times higher when compared to Digital channels. We are unable to disclose additional information for confidentiality reasons.

The dataset provided to us by our partners for this study was of exceptionally high quality, with an extremely low percentage of missing values, approximately only 0.0001%. This minimal amount of missing data suggested that the dataset was largely complete and reliable for analysis. For the few instances where missing values were present in certain categorical columns, we employed mode imputation as the method for filling these gaps. This approach was chosen with the assumption that substituting missing data with the mode would not cause any substantial distortion in the overall distribution of the data. Additionally, we applied label encoding to transform the categorical variables into numerical representations.

To address the challenge of class imbalance in the dataset, we employed a weighted training approach. Class imbalance in binary classification occurs when the number of observations in one class (the majority class) is significantly higher than the number of observations in another class (the minority class). Class imbalanced data can bias the model towards the majority class, thereby often leading to overfitting to majority class and underfitting to the minority class. This imbalance can significantly compromise model performance. Although oversampling techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE), are commonly used to balance class distributions by generating synthetic samples of the minority class, we chose not to implement this method. Our decision stemmed from the risk that synthetic data generation could produce unrealistic and invalid data points. E.g., creation of unrealistic categorical combinations, such as incorrect postal codes, which do not exist in the real world. The introduction of such synthetic instances could introduce biases into the model and negatively affect its ability to generalize to new, unseen data. In contrast, the weighted training method provides a more controlled and reliable

alternative. This approach involves assigning higher weights to the minority class data points during training, ensuring that the model gives greater emphasis to the minority class. By doing so, it effectively mitigates the impact of the class imbalance, allowing the model to learn without the risks associated with generating artificial data. weight w_j for a particular class j in a weighted classification is defined as:

$$w_j = \frac{n_{\text{samples}}}{n_{\text{classes}} \times n_j}$$

where n_{samples} represents total observations, n_{classes} the number of target classes, and n_j the count of class j instances.

4 Machine Learning Models and Experiments

Drawing insights from previous research, the study investigates six popular machine learning methods to ensure a comprehensive evaluation. Namely *Logistic Regression (LR)*, *Random Forest (RF)*, *Bagging of Decision Trees (BG)*, *XGBoost (XGB)* and *CatBoost (CB)*, *Shallow Neural Network (NN)* [4]. RF and LR, in addition to being robust to noise, have been extensively employed in past research in classification tasks. Furthermore, LR is widely used for its simplicity, interpretability, and effectiveness in binary classification problems. BG and XGB are well known for enhancing predictive performance while avoiding overfitting and improving generalization through bootstrap aggregation and regularised gradient trees [3, 5]. CB is optimized and well suited for categorical data, characteristic of features in the insurance domain [8]. Finally, we implement NN to capture complex non-linear relationships.

4.1 Model Training and Evaluation

The experimental design employs rigorous validation protocols beginning with a stratified 80:20 train-test split, preserving original class distributions to prevent evaluation bias. Hyperparameter optimization is conducted via exhaustive grid search (GridSearchCV) across predefined parameter spaces, with model selection finalized through 5-fold nested cross-validation during evaluation phases.

Performance Metrics Model performance is evaluated using three metrics that address class imbalance: Balanced Accuracy (BA), F1 score, and ROC-AUC. These metrics are defined below:

Balanced Accuracy(BA): Unlike traditional accuracy, which can be misleading in the presence of imbalanced data, balanced accuracy addresses this issue by considering both the majority and minority classes. In cases of class imbalance, a model that simply predicts the majority class can achieve a high accuracy score, despite performing poorly on the minority class. Balanced accuracy, however, ensures that the model's performance is evaluated on both classes equally, providing a more accurate reflection of its ability to classify instances

from both the majority and minority classes. By giving equal weight to the performance on each class, balanced accuracy prevents the model from being biased toward the majority class, making it a more reliable evaluation metric in situations with imbalanced data.

$$\text{BA} = \frac{1}{2}\left(\frac{\text{True Positive}}{\text{Positives}} + \frac{\text{True Negatives}}{\text{Negatives}}\right) \tag{1}$$

F1 score: It is defined as the harmonic mean of precision and recall, providing a single metric that balances these two often conflicting aspects of model performance. Precision is defined as the ratio of true positive predictions among all instances classified as positive. Recall quantifies the ratio of actual positive instances that the model correctly identifies.

$$\text{F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2}$$

Receiver Operating Characteristic - Area Under the Curve(ROC-AUC): It is a metric that evaluates overall ranking capability across classification thresholds. The metric is employed to evaluate the performance of classification models across different thresholds. It captures the trade-off between the true positive rate and false positive rate at various decision thresholds. The area under the ROC curve (AUC) quantifies the model's ability to discriminate between the positive and negative classes, with values closer to 1 indicating better performance.

5 Results and Discussion

The experimental results, summarized in Table 1, provide insights into the performance of various machine learning models across two distinct datasets.

5.1 Web Dataset Performance

Across all models BA ranged from 80.64% to 82.2%, while the F1 score varied from 85.9% to 87.9%. Among the models, CB achieved the highest BA of 82.2%, slightly outperforming other models. This was followed by NN with a BA of 82.1% and then XGB with BA of 82.01%. Although CB achieved the best performance in BA, NN demonstrated a slight advantage in terms of F1 score (87.9% vs. 87.7%), which indicates that NN offers a somewhat better precision-recall trade-off. This may imply that while CB maximized overall classification accuracy, NN provided a more balanced performance by minimizing false positives and false negatives. LR, despite its simplicity, performed adequately well (80.64% BA, 87.5% F1), which suggests that the Web dataset is close to being linearly separable, as shown in the Receiver operating curve (ROC) of Fig.1.

Table 1: Performance Metrics of Machine Learning Models on Web and Physical Datasets

Model	**Dataset**	**BA**	**F1**
LR	Web	80.64	87.5
RF	Web	81.22	87.4
BG	Web	81.65	85.9
XGB	Web	82.01	87.1
CB	Web	**82.20**	87.7
NN	Web	82.10	**87.9**
LR	Phy	71.66	71.9
RF	Phy	72.61	72.7
BG	Phy	72.34	72.1
XGB	Phy	72.51	72.6
CB	Phy	**73.66**	**73.6**
NN	Phy	72.70	72.5

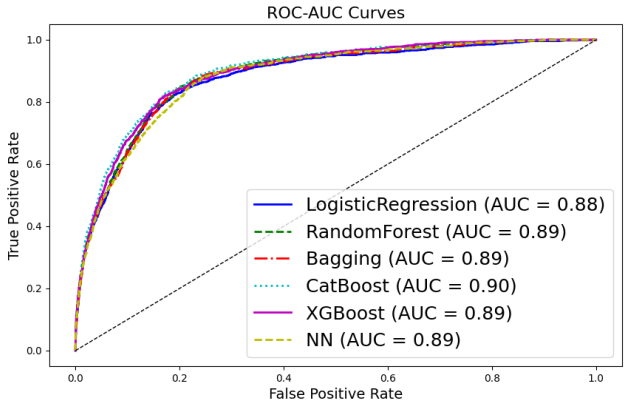

Fig. 1: ROC-AUC curve for Phy dataset

5.2 Physical Dataset Performance

The Physical dataset presented a comparatively challenging classification task, with all models showing lower performance compared to the Web dataset. BA ranged from 71.66% to 73.66%, while F1 scores were between 71.9% and 73.6%. CB once again emerged as the top performer, achieving the highest BA (73.66%) and F1 score (73.6%). The performance gap between the best and worst models was notably smaller in the Physical dataset (approximately 2% in BA) compared to the Web dataset (approximately 1.5%). This suggests that the Physical dataset may contain more complex or noisy patterns that are equally challenging for all model architectures, as shown in the ROC of Fig.2.

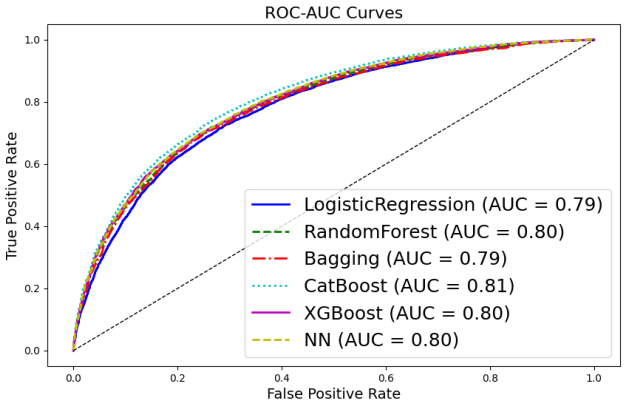

Fig. 2: ROC-AUC curve for Web dataset

5.3 Model Performance Overview

Across both datasets, CB emerged as the best-performing model, demonstrating superior predictive capabilities. CB achieved the highest BA and F1 scores in both datasets, underscoring its strong performance in both tasks. NN consistently performed a close second, highlighting the effectiveness of both gradient boosting and deep learning approaches for this task. When comparing different model types, Gradient boosting models (CB, XGB) outperformed Random Forest and Bagging. The likely reason for this is due to their iterative error correction process employed by gradient boosting models, which allows them to learn from their mistakes in successive iterations. Note that the relative performance of models remained fairly consistent across both datasets, with gradient boosting methods (CB and XGB) generally outperforming other approaches. This consistency suggests that these models could be preferential choices for similar classification tasks in the future. Our models achieved a reasonable performance score, with metrics aligned with results provided by the partner's internal models.

5.4 Impact of Sales Channel on Conversion Prediction and Customer Behavior

A critical aspect of this study was the observed differences in conversion rates and model performance between physical and digital sales channels. The significant performance disparity between physical and digital channels (approximately 10% in BA and 15% in F1 score) underscores the importance of channel-specific strategies in both sales approaches and modeling techniques.

The performance differences between physical and digital channels suggest distinct behavioral trends when customers are approached via different mediums. The higher success rate in physical sales might indicate that customers

value personal interactions and expert guidance when making insurance decisions. The lower conversion rate in digital channels suggests that online customers may require more personalized approaches or additional incentives to complete purchases. This could include strategies such as targeted offers or tailored recommendations to bridge the gap and improve conversion rates. These observations highlight the potential for tailored, channel-specific strategies to improve conversion rates, particularly in the digital sphere.

5.5 Conclusion

This study provides valuable insights into the performance of various machine learning models applied to customer conversion prediction, particularly focusing on the differences between physical and digital sales channels. Several limitations need to be addressed in future research. Further investigation is needed to understand the underlying causes of the performance gap between the Web and Physical datasets. Understanding these differences in more detail could guide future model refinements and help bridge the performance gap. Additionally, investigating the uncertainty within trained models would not only improve the reliability of predictions but also foster greater trust in machine learning-based decision-making. A critical observation from our research is the substantial performance disparity between the two sales channels, which emphasizes the importance of considering the unique characteristics of each channel when selecting models and performance expectations for real-world applications. Our study demonstrates the better performance of gradient boosting models, particularly CB, across both Web and Phy datasets in our classification task.

Bibliography

[1] Bag, S., Srivastava, G., Bashir, M.M.A., Kumari, S., Giannakis, M., Chowdhury, A.H.: Journey of customers in this digital era: Understanding the role of artificial intelligence technologies in user engagement and conversion. Benchmarking: An International Journal **29**(7), 2074–2098 (2022). https://doi.org/10.1108/BIJ-07-2021-0415

[2] Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition **30**(7), 1145–1159 (1997). https://doi.org/https://doi.org/10.1016/S0031-3203(96)00142-2

[3] Breiman, L.: Bagging predictors. Machine learning **24**, 123–140 (1996). https://doi.org/10.1007/BF00058655

[4] Chen, S., Xu, Z., Xu, D., Gou, X.: Customer purchase prediction in b2c e-business: A systematic review and future research agenda. Expert Systems with Applications p. 124261 (2024). https://doi.org/10.1016/j.eswa.2024.124261

[5] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785, `https://doi.org/10.1145/2939672.2939785`

[6] Jiangtao, Q., Zhangxi, L., Yinghong, L.: Predicting customer purchase behavior in the e-commerce context. Electronic Commerce Research **15**, 427 – 452 (2015). https://doi.org/10.1007/s10660-015-9191-6

[7] Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., Haltmeier, M.: A machine learning framework for customer purchase prediction in the non-contractual setting. Eur. J. Oper. Res. **281**, 588–596 (2020). https://doi.org/10.1016/J.EJOR.2018.04.034

[8] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A.: Catboost: unbiased boosting with categorical features. Advances in neural information processing systems **31** (2018). https://doi.org/10.5555/3327757.3327770

[9] Semih, B., Emin, A., Ahmet, B.: Forecasting conversion rate for real time cpc bidding with target roas. IEEE Access **11**, 134908–134916 (2023). https://doi.org/10.1109/ACCESS.2023.3338022

[10] Soni, V.: Role of temporal, demographic, and behavioral factors in customer conversion through dynamic creative optimization in the consumer-packaged goods setting. International Journal of Business Intelligence and Big Data Analytics **5**(1), 46–56 (2022)

[11] Tang, L., Wang, X., Kim, E.: Predicting conversion rates in online hotel bookings with customer reviews. Journal of Theoretical and Applied Electronic Commerce Research **17**(4), 1264–1278 (2022). https://doi.org/10.3390/jtaer17040064

[12] Tokuç, A.A., Dağ, T.: Customer purchase intent prediction using feature aggregation on e-commerce clickstream data. 2024 8th International Arti-

ficial Intelligence and Data Processing Symposium (IDAP) pp. 1–5 (2024). https://doi.org/10.1109/IDAP64064.2024.10711144

[13] Wendi, J., Xiaoling, W., Feida, Z.: Time-aware conversion prediction. Frontiers of Computer Science **11**, 702–716 (2017). https://doi.org/10.1007/s11704-016-5546-y

[14] Yeo, J., Hwang, S.w., kim, s., Koh, E., Lipka, N.: Conversion prediction from clickstream: Modeling market prediction and customer predictability. IEEE Transactions on Knowledge and Data Engineering **32**(2), 246–259 (2020). https://doi.org/10.1109/TKDE.2018.2884467