Pitchbook Data Science Project

By: Jonathan Campbell

Last Updated: 2023-11-17

Introduction

In answering the questions provided, we're diving into a detailed analysis with a key goal: figuring out which opportunities are likely to turn into actual sales. To further that goal, we will be building a simple model that predicts the chance of a sales opportunity becoming successful. This model relies on understanding how different factors, like the account and sales rep details, impact the likelihood of a sale.

Our aim is not just to answer the questions but also to provide practical insights that can help the business make informed decisions about sales opportunities. We're here to bring clarity to areas of potential, highlight any challenges, and suggest smart strategies for improving sales outcomes.

Analysis

The dataset under analysis comprises mostly categorical data related to sales opportunities, encompassing details about accounts and sales representatives. Each entry in the dataset provides a snapshot of specific instances where sales were either won or lost. Key attributes include account-related tags, sales representative profiles and their locations, and the outcomes of the respective sales opportunities. This dataset allows us to explore the relationships between various factors and the success or failure of sales efforts by noting the timestamps of each of the stages reached throughout the process.

The data gave us a couple of challenges. First, there were a lot of missing details about accounts, leaving many accounts without information. Second, there's an uneven distribution in the outcomes – we've got way more lost opportunities than ones that turned into sales. Lastly, we noticed that the sales that did succeed went through different stages way more often than the ones that didn't. This means that these values will likely be highly predictive of a sale.

In the preprocessing phase, I undertook several transformations to enhance the dataset. First and foremost, all categorical features were converted into binary variables through a process known as one-hot encoding. This conversion allows the model to effectively interpret and utilize categorical information, breaking down each category into individual binary features. This transformation ensures that the model can incorporate non-numeric data, such as account and sales representative office locations, into its learning process.

I leveraged the dates associated with sales stages by translating the dates into additional binary variables. These binary variables represent whether the opportunity ever reach a certain stage in the sales process. To capture the temporal aspect of the data, I included the time from creation to the first event reach and the time from the last event reached to the opportunity being closed.

To address the class imbalance inherent in the dataset, I employed a stratified method for the train-test split to ensure that the proportion of converted and lost opportunities remains consistent in both the training and testing sets. This safeguards against potential biases that could arise from an uneven distribution of classes and enhances the model's ability to generalize across different outcomes.

In building the predictive model, I opted for a logistic regression framework with Lasso regularization. Lasso regularization introduces a penalty term that encourages sparsity in the model coefficients, effectively performing feature selection. This is particularly beneficial in our context, where the dataset exhibits high dimensionality. By encouraging sparsity, Lasso helps in identifying the most influential features for predicting sales conversion, mitigating the impact of less informative variables. This regularization technique contributes to the model's interpretability, aids in handling multicollinearity, and enhances its generalization performance, especially in the context of imbalanced datasets.

In employing a logistic regression, several considerations arise, particularly given the predominantly categorical nature of the data. Logistic regression's inherent difficulty in handling non-linear relationships can limit its capacity to capture nuanced patterns, and as such, alternative models may be warranted in situations where non-linearities play a substantial role. Additionally, logistic regression may pose challenges in accommodating intricate feature engineering, such as interactions between variables, which can be crucial for capturing the complexity of relationships in categorical data. As we interpret the results and insights from our logistic regression model, it is imperative to be cognizant of these limitations and consider alternative modeling approaches.

Model Outcomes

The model's performance, as indicated by the confusion matrix and accompanying metrics, showcases a better predictive ability than average. The confusion matrix reveals a total of 4580 true negatives, 186 true positives, 52 false positives, and 179 false negatives. The precision of 0.78 signifies a commendable proportion of correctly predicted conversions among instances predicted as positive, highlighting the model's reliability in avoiding false positives. While the recall of 0.51 indicates that the model successfully identified around half of the actual conversions, there is room for improvement in capturing more positive instances. The overall accuracy, however, stands at 0.95, indicating the model's proficiency in making correct predictions across both classes.

Actual \\ Predicted	Negative	Positive

Negative	4580	52
Positive	179	186

Precision: 0.78
Recall: 0.51
Accuracy: 0.95

Insights

Examining the trends in the logistic regression coefficients, certain patterns emerge regarding the likelihood of sales conversion across different business types. Commercial businesses exhibit a negative log odds, suggesting a decreased likelihood of conversion for this category. In contrast, corporate businesses show positive log odds, indicating an increased probability of successful sales outcomes. Notably, the healthcare and pharmaceutical sector also displays positive log odds, pointing towards a higher likelihood of sales conversion. Similarly, businesses associated with capital lenders and venture capital demonstrate positive log odds, implying an elevated probability of successful conversions for opportunities linked to these entities. These trends shed light on the diverse impact that various business types can have on the predictive model, offering valuable insights for strategic decision-making in sales and business development efforts.

In terms of Pitchbook’s business, this might mean that the business is strategically situated towards businesses focusing on long term relationships, either with customers or with other businesses. These businesses also seem to offer some type of specialized service. The specialized nature of these offerings may contribute to the positive impact on the likelihood of conversion, as clients may value specific expertise and tailored data.

The analysis reveals a noteworthy trend indicating that progression through various sales stages is positively predictive of successful outcomes. It's crucial to acknowledge and scrutinize the inherent bias associated with this observation. The model's reliance on reaching different checkmarks or milestones as a proxy for success introduces a potential circularity. The criteria for success, in this case, are intertwined with the very stages used to predict success. While the positive predictiveness of reaching stages is informative, it underscores the importance of interpreting results cautiously and considering the nuanced nature of the dataset. A deeper examination is warranted to disentangle the impact of reaching stages from other factors and to ensure that the predictive power of the model is grounded in meaningful, rather than tautological, insights.

Appendix

Variable	Log Odds
OFFICE_London	0.360574
OFFICE_New York	-0.111547
PRIMARY_TYPE_Company	-0.175869
PRIMARY_TYPE_Limited Partner	0.037709
PRIMARY_TYPE_Service Provider	-0.207039
SECONDARY_TYPE_Accelerator/Incubator	-0.370799
SECONDARY_TYPE_Accounting/Auditor	-0.343186
SECONDARY_TYPE_Angel (individual)	-1.220555
SECONDARY_TYPE_Angel Group	-1.517484
SECONDARY_TYPE_Asset Manager	-0.299674
SECONDARY_TYPE_Capital Markets/Institutions	-0.131986
SECONDARY_TYPE_Commercial Banks	-0.518465
SECONDARY_TYPE_Commercial Products	-0.336574
SECONDARY_TYPE_Commercial Services	-0.483137
SECONDARY_TYPE_Commercial Transportation	-0.504886
SECONDARY_TYPE_Consumer Non-Durables	0.696377
SECONDARY_TYPE_Containers and Packaging	-0.202037
SECONDARY_TYPE_Corporate Development	0.895987
SECONDARY_TYPE_Corporate Pension	0.174224
SECONDARY_TYPE_Corporate Venture Capital	0.648384
SECONDARY_TYPE_Corporation	-0.167123
SECONDARY_TYPE_Energy Services	-0.097122
SECONDARY_TYPE_Exploration, Production and Ref...	-0.310292
SECONDARY_TYPE_Family Office	0.457317
SECONDARY_TYPE_Family Office (Single)	0.025507
SECONDARY_TYPE_Financing Advisory	-0.112635
SECONDARY_TYPE_Funding/Crowdfunding Platform	0.020169
SECONDARY_TYPE_Fundless Sponsor	-0.248815
SECONDARY_TYPE_Government	1.047626
SECONDARY_TYPE_Growth/Expansion	0.309122
SECONDARY_TYPE_Healthcare Devices and	0.391425

Supplies	
SECONDARY_TYPE_Healthcare Services	0.414092
SECONDARY_TYPE_Healthcare Technology Systems	0.432383
SECONDARY_TYPE_Hedge Fund	0.087929
SECONDARY_TYPE_IT Services	-0.459268
SECONDARY_TYPE_Investment Bank	-0.046049
SECONDARY_TYPE_LP Consultants	0.094545
SECONDARY_TYPE_Law Firm	-0.725894
SECONDARY_TYPE_Lender	0.441184
SECONDARY_TYPE_Lender/Debt Provider	0.131711
SECONDARY_TYPE_Limited Partner	0.780195
SECONDARY_TYPE_Management Consultants	-0.692664
SECONDARY_TYPE_Media	0.577986
SECONDARY_TYPE_Merchant Banking Firm	0.202261
SECONDARY_TYPE_Other Business Products and Ser...	-0.204409
SECONDARY_TYPE_Other Information Technology	0.523446
SECONDARY_TYPE_Other Private Equity	0.175771
SECONDARY_TYPE_PE/Buyout	0.045961
SECONDARY_TYPE_PR Firm	0.288403
SECONDARY_TYPE_Pharmaceuticals and Biotechnology	0.145714
SECONDARY_TYPE_Private Equity	-0.192717
SECONDARY_TYPE_Real Estate	-0.618684
SECONDARY_TYPE_Recruiting Firm	0.497836
SECONDARY_TYPE_Restaurants, Hotels and Leisure	0.18362
SECONDARY_TYPE_SPAC	1.362075
SECONDARY_TYPE_Services (Non-Financial)	-0.882541
SECONDARY_TYPE_Software	0.186639
SECONDARY_TYPE_Valuation Firm	-0.238643
SECONDARY_TYPE_Venture Capital	0.3676
SECONDARY_TYPE_Wealth Management Firm	-0.442626
SECONDARY_TYPE_nan	-5.287682

DATE_OF_QUALIFYING	0.217213
DATE_OF_EVALUATION	2.005927
DATE_OF_PROCUREMENT_NEGOTIATIONS	3.499084
DATE_OF_VERBAL_PENDING	4.556524
OPPORTUNITY_OPEN_DAYS	-0.527554
FIRST_EVENT_DAYS	0.249234
LAST_EVENT_DAYS	0.106807