Shiyu Ma

Professor Richard Wesel

Engineering 195

24 August 2022

Final Report

My major intern project at Thumbtack is called the 'Lead Score Model'. The goal is to predict conversion probabilities of potential pros, who are service providers in the two-sided market created by Thumbtack. By using this model, our team hopes to improve the efficiency of the sales team when they reach out to potential pros in the lack of supply local market. Meanwhile, we hope to identify characteristics of high potential pros and gain a better understanding of pros acquisition.

The biggest challenge I faced is the imbalance data issue. With an overall conversion rate of 7%, the model could get an accuracy of 93% by predicting all samples as negative. However, this leads to a useless model that fails to distinguish positive cases from the rest. To solve this issue, I decided to use recall as the performance metric after confirming our priority of catching all possible conversions. This process also exposed me to differences across other classification metrics such as precision, F1 score, and AUC. Furthermore, to enforce the model focusing on positive cases, I researched 4 possible sampling methods and applied 2 of them to the later modeling part. This experience allowed me to learn the pros and cons of Stratified Sampling, SMOTE Oversampling, and Cost-Sensitive Learning while understanding how to choose between them based on data structure and the specific model we use.

My major accomplishment in this project is feature selection. There are 62 features available in the dataset. Such a large number of features hurt interpretability and violate feature independence assumption of many models, so feature engineering and selection are required. To decide on the method, I explored feature distribution and found them to be highly skewed. With such data, model with strict assumptions is unable to give reliable feature importance. Therefore, the random forest becomes my feature selection tool that is robust to outliers and data distribution. I wrote my own function to generate metrics and features' ranks from multiple rounds of sampling. My function effectively eliminates subtle model differences due to randomness and gives lists of least useful features to suggest dropping. Combined with my business intuition on features, I selected 4 features out of 62 while achieving a higher recall than before.

Though the model has a satisfactory recall, there are spaces to improve in precision. I believe further feature engineering and mining are promising solutions. Many of the features are the same metrics across different platforms, so I tried multiple ways to aggregate and regroup them into new features. However, such an effort of catching interaction between platforms does not improve our performance significantly. Also, I learned a statistical method that takes the number of reviews behind ratings into consideration. Though it seems to correct bias in rating, the resulting feature fails to boost our performance. In the future, we may enrich our dataset further and add new features based on sales experts' insights. Although my attempts of feature engineering did not work out well, this experience taught me how important domain knowledge could be as a data analyst.

In general, I got a deeper understanding of classification problems, sampling methods, and feature engineering from this project. In particular, I learned how to deal with imperfect datasets and practical constraints. Though real-world modeling is more challenging, the business motivation behind makes the model more impactful and worthful.

Besides the lead score model, I also helped with conducting AB testing on a marketing campaign. This side-project leads me to learn several new skills, including Tableau, Mode, and Power Analysis. Also, I got exposure to experimental design in such a two-sided market. These are exactly what I expected to learn before the internship and I am so glad that I got the chance to learn all of them.

For the next step of my career growth, I hope to explore large-scale machine learning such as NLP, recommendation algorithms and search ranking models if I end up pursuing a master's degree. Gaining insights from models is an interesting task for me, so I prefer to learn more interpretable models over black boxes. However, if I end up getting into the industry right after my bachelor's degree, I would build myself a mindset on how to solve business problems by data and accumulate analytics methods I could apply in various cases. Both tracks are equally attractive to me and I will see how future opportunity leads me to the next step.

Manager Analytics Thumbtack