Writing Sample
Ratul Esrar

The Los Angeles County Department of Mental Health (DMH) hired a consultancy to develop two machine learning (ML) models that predict whether or not a DMH client will enter an emergency shelter in the 12 months following their first visit to a DMH clinic. DMH estimates they are able to serve a maximum of 500 people in a given year through a targeted intervention strategy. ***This memo analyzes the ML models using evaluation frameworks for supervised learning, and recommends which model to deploy for a targeted homelessness intervention system.***

To Los Angeles County Department of Mental Health (DMH):

Executive Summary
In this memo, I evaluate the two candidate models provided to DMH and recommend that DMH use Model 2 (M2) to best inform their intervention strategy in which they target a subset of the 500 highest-risk clients for proactive outreach by a specialist team of caseworkers. Although the costs involved are high, this strategy is likely to have the highest impact and ROI. Proactive outreach will not only benefit these high-risk and vulnerable clients, but also may have positive externalities on other individuals and organizations.

Model Analysis
I analyzed a dataset of 52,843 clients who visited DMH for the first time in the year 2018 and found that 6,439 (12.2%) individuals entered a homeless shelter in the 12 months following their first DMH visit date. The data also included information about arrest records, which were likely incorporated as inputs in the models. The arrests data are not balanced in terms of race, which can introduce bias to the models. Since the provided dataset used for model validation has over 50k records, roughly the top 1% yielded the 500 highest-risk individuals for intervention. An evaluation metric at 1% describes how the model performs on only that 1% of cases. Since resources are limited, it is best to optimize for precision at 1% because this metric minimizes the rate of false positives. Table 1 (see below) describes several model evaluation metrics, including accuracy, precision, recall, and f1-score.

Table 1: Evaluation Metrics by Model

Table 1: Evaluation Metrics by Model at Top 1%				
	Accuracy	**Precision**	**Recall**	**F1-Score**
Model 1	88.4%	81.8%	6.4%	11.8%
Model 2	88.5%	87%	6.8%	12.5%

Using the top 1% as a threshold, I calculated several model evaluation metrics to compare the two models. These are summarized in Table 1 and below is each column's interpretation:

- Accuracy is how often the model correctly predicted the outcome of entering a homeless shelter within 12 months of an individual's first DMH visit.

- Precision describes the proportion of positive predictions by the model where the individual actually entered a homeless shelter.
- Recall describes the proportion of actual homeless outcomes that the model identified correctly.
- F1-Score is the harmonic mean between precision and recall, and can be used to evaluate models and strike balance between precision and recall.

Figures 1, 2, and 3 in the appendix visualize the predicted probabilities of the candidate models and include additional commentary around model selection. The models provided by the consulting firm had very different ranges of prediction values for the homeless shelter outcome. Figure 1 shows the distribution of predicted probabilities calculated by the two models. Model 1 (M1) suggested over 20% of individuals have a 0% chance of entering a homeless shelter in the 12 months following their first DMH visit. M1 also assigned 80% of individuals less than a 20% chance of entering a homeless shelter. M1's positive predictions are almost uniformly distributed. On the other hand, Model 2's predictions ranged from 44% to 56% and had much more overlap between homeless-positive and homeless-negative groups. Figures 2 and 3 display the distribution of predicted probabilities from M1 and M2, and include further analysis of the model predictions.

These considerations also played a role in recommending an intervention strategy. Translating the model predictions into scaled risk-scores is complicated and difficult to do without insight into how the models made their predictions. As a result, I used the precision at 1% to recommend M2 as the best model given the intervention strategy.

Recommendations

For the 500 highest-risk individuals, both models accurately predicted whether or not they entered a homeless shelter around 88% of the time. This is over 7 times more accurate than if individuals were selected at random for outreach since the observed homeless rate among clients was 12.2%. Model 2 had the higher precision at 1%, which means that it correctly predicted a true positive homeless outcome for 435 out of 500 people. M2 also had a higher f1-score than M1, which suggests that M2 is the optimal model to deploy if DMH is able to expand the outreach program and target a larger number of at-risk individuals.

In doing so, it is important to consider potential bias in the data used to train the model as well as additional ethical implications of using algorithms to inform intervention systems. Models can inherit implicit biases when they are trained using race/ethnicity, gender, and arrest information from institutions that have perpetuated systemic racism. For example, black people are overrepresented in the provided arrests dataset. Of the 6,984 clients who had an arrest prior to their first DMH visit, 47% are listed as black. Based on census estimates for LA County as of July 2019, an estimated 9% of the county is black.[1] This overrepresentation might be reflected in the model, which can reinforce pre-existing racial and social biases. When training and

[1] https://www.census.gov/quickfacts/losangelescountycalifornia

evaluating any predictive model, it is critical to be aware of these types of imbalances in the data and adjust for these by choosing a representative training dataset. In addition, transparency around algorithm-assisted decision making processes is important to prevent harm to the people that DMH serves.

Conclusion

I recommend DMH deploys Model 2, as this minimizes the rate of false positives when intervening on the 500 highest-risk individuals. This allows DMH to proactively provide critical services to the most vulnerable clients, and this model is likely to perform better if DMH procures more resources and expands the intervention program. Prior to deploying the model, the consulting firm and DMH should analyze the training data for potential bias and imbalance and correct for these using other statistical techniques when possible.

Appendix

Figure 1: Distribution of Prediction Probabilities by Model

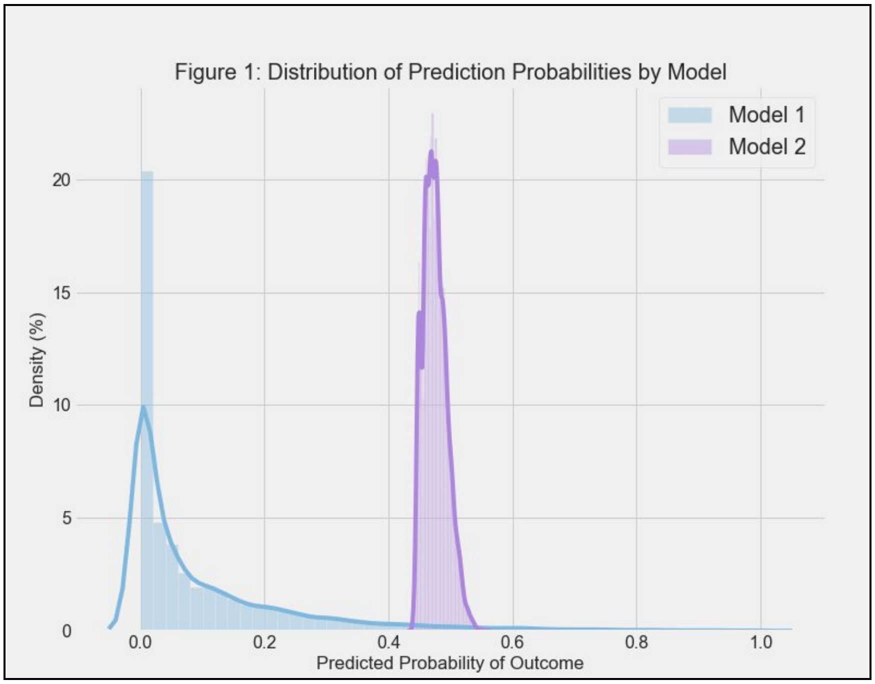

Figure 1 displays the distribution of predicted probabilities by each model. M2's predictions are unimodal and range from 44% to 56% while M1's predictions are right-skewed and range from 0 to 1.

Figure 2: Distribution of M1 Prediction Probabilities based on Outcome

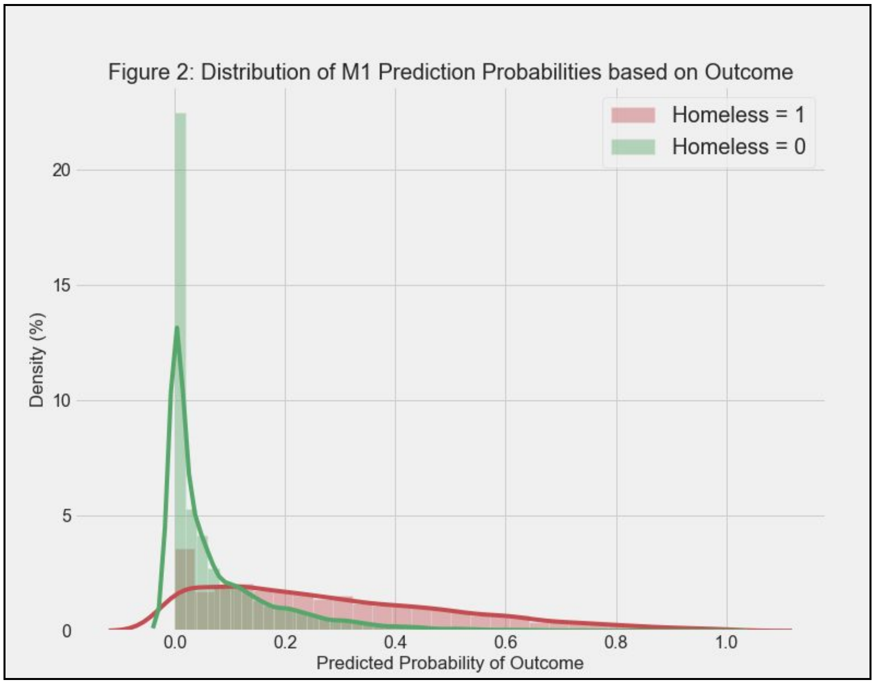

Figure 2 displays the distribution of predicted probabilities from M1 split out by whether or not an individual actually entered a homeless shelter. Since the distribution is skewed, the average predictions could be heavily influenced by outliers. Using medians, M1 did assign individuals who entered a homeless shelter higher predicted values with a median of 24.1% chance compared to the median of 3.9% for those who did not enter a shelter.

Figure 3: Distribution of M2 Prediction Probabilities based on Outcome

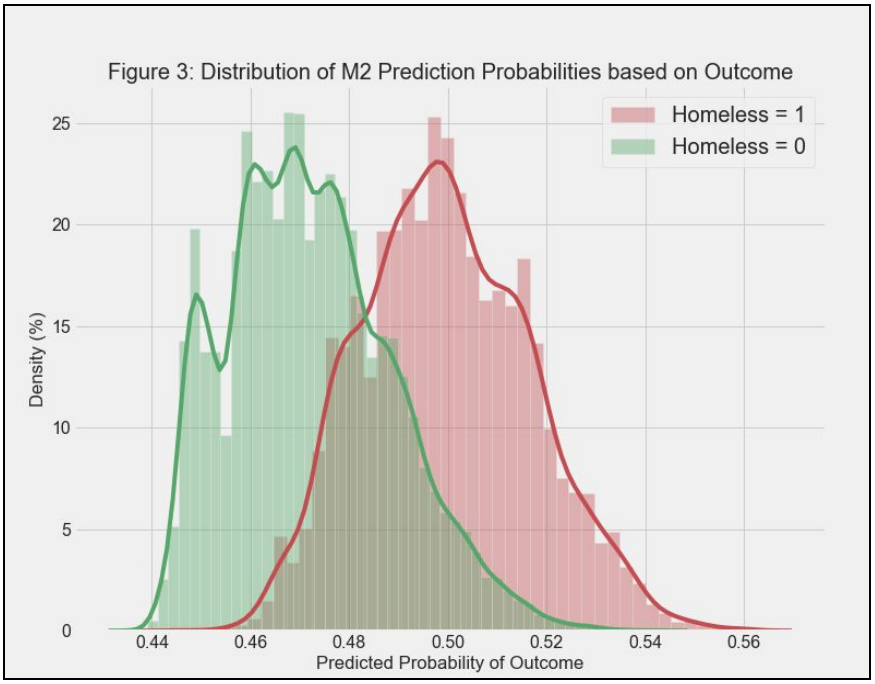

Figure 3 displays the distribution of predicted probabilities from M2 split out by whether or not an individual actually entered a homeless shelter. M2 also assigned individuals who entered a homeless shelter higher predicted values with an average of a 50% chance compared to the average of 47.2% for those who did not enter a shelter. There is much more overlap in prediction values between the homeless and not homeless groups for M2.