**Predicting Homeless Shelter Utilization in Washington DC**

The purpose of this project is to help the DC department of housing services, in collaboration with DataKind DC, better predict walk-in, homeless shelter need. For context, Washington DC is one of a limited number of US cities that believe in the Right-To-Shelter, and as such it guarantees walk-in (low-barrier) shelter for anyone who needs it. Of the roughly 6000 homeless residents, about one third typically utilize the walk-in shelter services during the winter months. These walk-in shelter sites are comprised of a network of stable year-round shelters, winter-specific sites, and overflow shelter that are opened as needed (at variable and typically undisclosed locations). Many shelters are male or female specific; buses are utilized to transport residents to shelters as needed. Additionally, on nights when the weather conditions are deemed to pose a significant hypothermia threat, the police will go into the community and attempt to convince as many homeless people as possible to enter the shelters. For this project only the most unstable form of shelter support, that of individual walk-in shelter, will be considered, rather than the more easily planned for housing support services given to family housing, transitional housing, or rental assistance.

The code included in this Git repository is my contribution to the project. I have received CSV data regarding male and female nightly shelter usage, per shelter, since 2007 from DataKind DC. Although I’m unable to share this data publicly, but I have included all my processing and machine learning code here (in python and Jupyter notebook versions), and you can view the results of my modeling at <https://shelter-usage-predictions.herokuapp.com>. During processing I discovered one shelter that had an unusually high activity reported in 2019 (and it was subsequently agreed up that it should be removed), plus another a second shelter that housing families (which has also excluded from the individual count). The women’s shelter numbers display steep jumps in 2015 and 2016, so for modeling purposes I am restricting the training set to 2016 and onward. Even so, the women’s shelter predictions remain less reliable than the male shelter or total shelter predictions due to a lack of multi-year, consistent data.

With shelter data, I’ve combined weather information, including minimum and maximum daily temperature, previous day’s minimum temperature, wind chill, rain precipitation and snow accumulation, obtained from Visual Crossing Weather. I’ve included general DC resident yearly census population numbers, and the homeless individual population as surveyed once a year in January as part of the Point-In-Time count of homeless people on the street plus the individuals utilizing the walk-in shelter services that night (consistent forms of this data can be found from 2011 onward). For modeling purposes, sine and cosine forms of annual periodicity were added to help capture yearly trends, and day of the week was also included (where shelter usage peaks mid-workweek). Lastly, DC MetroBus ridership was included as a proxy for predicting the reduction in congregate setting activity with the onset of Covid-19 in early 2020. Even so, the model could not predict that women were about 20% more likely not to utilize walk-in shelters compared to their male counterparts during the 2020-2021 season coinciding with the Covid-19 pandemic. Moreover, the introduction of individually assigned hotel sites for the medically vulnerable population in 2020 altered the yearly trends further such that these people were far more likely to remain in their assigned sites than normal, especially during the summer months.

For the machine learning, the variables discussed above were either one-hot-encoded or fed into a linear regression model, which was then combined using random forest regression of depth 7. I noticed from multi-year plots of temperature versus shelter usage that decreasing minimum daily temperatures below about 26 F to 30 F did not seem to result in a further increase in shelter usage. Thus, the model performs better when the data are broken up into low-temperature and mid-temperature groups at the 30 F boundary and trained independently.

The results of my model can be accessed at <https://shelter-usage-predictions.herokuapp.com>, where one can interactively query the population type, data at which the model training ends, the timespan of the prediction, and compare the results to known shelter usage (when applicable). The year-long predictions are useful for long-term planning where August is typically the time when decisions regarding how much shelter space to contract for the upcoming hypothermia season must be made. On this level, the full model (with knowledge of the daily weather) typically does somewhat better a naïve model that simply scales the previous year’s shelter usage by homeless population. However, for 10-day predictions initiated anytime during the hypothermia season, the full model significantly outperforms the naïve model. This could be highly relevant for refining how much overflow shelter space to activate based on an upcoming ten-day forecast.

I hope you enjoy learning about this project, and please feel free to contact me with any questions or comments at Kathryn.Krycka@mg.thedataincubator.com.

Sincerely,

Kathryn Krycka