

1 Two-Step Light Gradient Boosted Model to identify human West Nile

2 Virus infection risk factor in Chicago

3 Guangya (Wayne) Wan^{1,2}, Joshua Allen¹, Weihao Ge¹, Shubham Rawlani^{1,3}, John Uelmen⁴, Liudmila

4 Sergeevna Mainzer^{1,5}, Rebecca Lee Smith^{1,4,5}

5 1. National Center for Supercomputing Applications, University of Illinois, Urbana-Champaign 2.

6 University of Illinois, Department of Statistics, 3. University of Illinois, Information School, 4. University

7 of Illinois, Department of Pathobiology, University of Illinois, Urbana-Champaign, 5. Carl R. Woese

8 Institute for Genomic Biology, University of Illinois at Urbana-Champaign.

9 Abstract

10 West Nile virus (WNV), a flavivirus transmitted by mosquito bites, causes primarily mild symptoms but
11 can also be fatal. Therefore, predicting and controlling the spread of West Nile virus is essential for public
12 health in endemic areas. We hypothesized that socioeconomic factors may influence human risk from
13 WNV. We analyzed a list of weather, land use, mosquito surveillance, and socioeconomic variables for
14 predicting WNV cases in 1-km hexagonal grids across the Chicago metropolitan area. We used a two-
15 stage lightGBM approach to perform the analysis and found that hexagons with incomes above and below
16 the median are influenced by the same top characteristics. We found that weather factors and mosquito
17 infection rates were the strongest common factors. Land use and socioeconomic variables had relatively
18 small contributions in predicting WNV cases. The Light GBM handles unbalanced data sets well and
19 provides meaningful predictions of the risk of epidemic disease outbreaks.

20

21 Introduction

22 West Nile Virus (WNV) is a mosquito-borne flavivirus that has been circulating in the United States for
23 two decades, first appearing in New York in 1999 [1–3]. The disease is spread in an enzootic mosquito-
24 bird-mosquito circulation [4–7], and zoonotic transmission occurs when humans are bitten by a WNV-
25 positive mosquito [8]. Because there are no vaccines for WNV in humans, prediction of WNV-positive
26 mosquitoes is used to inform public health actions to clear mosquitoes in areas of high risk [9] and to
27 warn the general public of increased risk.

28 Efforts have been made to build predictive models of WNV spread [10]. Predicting human cases would
29 help to identify high-risk populations, and therefore enable protective measures. Paz [11] analyzed major
30 weather factors and found temperature and precipitation are associated with WNV human cases. A
31 temperature range of 10-35°C is advantageous for mosquito breeding activity. However, an association of

32 temperature with WNV infection risk is not always positive. Hahn et al.[12] performed a climate-region-
33 wise analysis and found that in most regions of the US, temperature above the local average increases
34 WNV risk, while in the western regions of the US, above-average temperature decreases WNV risk.
35 Shocket et al. [13] has identified the optimal temperature range for mosquitoes that vector WNV is
36 between 23-26°C. Precipitation and humidity have complex associations with mosquito population and
37 infection rate, as well. Interaction between temperature and precipitation also explains a significant part of
38 the WNV mosquito infection rate [14]. Poh et al. identified that temperature and rainfall increase
39 mosquito abundance [15]. In addition to temperature and precipitation, other factors such as humidity and
40 wind velocity affect mosquito abundance [16]. Peper et al. have studied WNV and mosquito surveillance
41 records from Lubbock, TX, and have found that the probability of mosquito infection depends on the
42 weather variables including the time in the year, wind, visibility, humidity, dew point, and the time lag of
43 these variables [17]. They also found that weather has a temporal autocorrelation, which brings lagging
44 effects into play [18,19]. DeFelice has discussed the lag in reporting of both mosquito infection and
45 human cases that reduces real-time WNV forecast accuracy and proposed recursive optimization and
46 Poisson process simulation for the retrospective forecast to solve the problem [20]. The landscape also
47 contributes to WNV risk. Studies have identified geological factors such as vegetation, urbanization,
48 mosquito breeding sites, and wetlands to be associated with WNV incidences [21–23]. Sánchez-Gómez et
49 al. have discussed how temperature and the presence of wetlands influence WNV circulation in vectors
50 and humans [21]. Hernandez et al. have identified weather, demographic, and controlling measurements
51 including temperature, precipitation, ethnicity, mosquito breeding sites, targeted prevention, and
52 education as key predictors [22]. Myer and Johnston have analyzed a 15-year span of data in Nassau
53 County, NY, and identified landscape factors including high normalized difference vegetation index
54 (NDVI), wetlands, and high urban development have a negative association with WNV incidences [23].
55 Farooq et.al. have estimated WNV expansion risk and found early spring weather, population, and
56 agriculture activities can be important factors for early warning systems to predict Europe WNV outbreak
57 [24]. Bassal et.al. investigated demographic disparities for WNV IgG levels in Israel and identified

58 different WNV seroprevalence among geographical regions. Bassal et. al. also discovered different
59 prevalence among racial groups, which have different socioeconomic status [25].

60 Linear regression and ensemble tree methods are the two most commonly used approaches for predicting
61 WNV incidence or mosquito populations. Hernandez et al. started with chi-squared tests to identify a list
62 of candidate factors and then used regression to find the strongest predictors [22]. Karki et al. used a
63 stepwise model selection procedure to automatically test all factors and find the strongest predictors [26].
64 However, the risk of WNV is not linear with the factors. Furthermore, linear models have high specificity
65 and perform best when there are no cases of viral infection, but have poor sensitivity when there are cases
66 (low recall). To address these two issues, ensemble methods, specifically light gradient boosting method
67 (GBM) approach [27], are used as our model in this paper. Light GBM is based on building an ensemble
68 of decision trees instead of a single model to make the prediction. Therefore, neither requires linearity in
69 the problem. However, light GBM is much faster to train and evaluate than other methods such as random
70 forest [28,29], has a generally lower bias, and thus will be our focus in this paper. We performed a two-
71 step light GBM approach as recommended for other ensemble tree methods [28,29]. In the first step, all
72 factors are included in the model. And then a second light GBM classification/regression is performed
73 based on the top factors selected by the first model [28].

74 We have hypothesized that, in addition to natural factors such as mosquito infection rate (MIR), weekly
75 temperature, temperature in January, and precipitation, social economics and land cover factors will also
76 be predictive factors for the WNV occurrences. We also hypothesized that natural factors might have
77 lagging effects. These effects, linear or not, can be detected by the light GBM approach and identify areas
78 at high risk of WNV cases and provide guidance for health intervention.

79

80 Methods

81 Data Set and pre-analysis

82 The dataset we used is described in more detail in Karki, et al. [26]. The dataset includes the number of
83 human disease cases from 2005-2016 in Cook and DuPage Counties, IL, as the dependent variable, and
84 several independent variables comprising weather, socioeconomic, land cover, and mosquito infection
85 rates (MIR). All variables were aggregated on a weekly temporal resolution and on a spatial grid of 1 km
86 wide hexagons for the study region.

87 The human disease data is described as a binary number that represents whether a case occurs in a
88 hexagon in a given week. We performed the two-sample Kolmogorov-Smirnov (KS) test [30] and the
89 two-step light GBM classification [27] to build the model to predict the human illness data and to derive
90 the illness probability from the model.

91 Weather variables include temperature and precipitation, as well as the lagged variables representing
92 temperature and precipitation 1 week, 2 weeks, 3 weeks, and 4 weeks before human case report date. The
93 original weather data was collected by PRISM [31], aggregated to census tract level, and mapped to
94 hexagons by Karki, et al. [26]

95 The land cover data include urban areas (developed open space, developed low intensity, developed
96 medium intensity, developed high intensity), forest (deciduous, evergreen, and mixed), barren land,
97 shrubs, grassland, pasture, cultivated crops, woody wetlands, herbaceous wetlands, and open water.
98 Karki, et al.[26] retrieved the land cover data from the 2016 National Land Cover Database (NLCD) [32]
99 and aggregated the percentage of different land covers in the hexagons.

100 For the socioeconomic data used by Karki, et al. [26], the 2016 census data from the US Census Bureau
101 [33] was applied across all years. The data were converted from the census tract level to the hexagon level

102 by assuming homogeneous socioeconomic status within each census tract. To determine the sensitivity of
 103 the socioeconomic data to annual changes, we replicated the mapping procedure with the 5-year rolling
 104 averages from 2010-2017 and performed the model analysis with both datasets (S1, S2). We found that
 105 the results are similar and the conclusions do not change; therefore, we will present the model built with
 106 the 2016 census data.

107 The variables we used are listed in Table 1 below.

108 **Table 1. List of variables involved in building the models. We have variables representing nature**
 109 **factors, land cover, and socioeconomic data.**

Notation	Variable	Type
mir_mean, mir_lag1-4	Mosquito Infection Rate (MIR) measured 0-4 weeks before human case report date	nature
preci, preci_lag1-4	Precipitation measured 0-4 weeks before human case report date	
tempc, temp_lag1-4	Temperature measured 0-4 weeks before human case report date	
tempJan	Temperature in January	
dospct	Proportion of developed open space	Land cover
dlipect	Proportion of developed low intensity	
dmipct	Proportion of developed medium intensity	
dhipct	Proportion of developed high intensity	

dfpct	Proportion of deciduous forests	
efpct	Proportion of evergreen forests	
mfpct	Proportion of mixed forests	
blpct	Proportion of barren land	
shrubspct	Proportion of shrubs	
glandpct	Proportion of grassland	
pasturepct	Proportion of pasture	
clpct	Proportion of cultivated land	
wwpct	Proportion of woody wetlands	
shwpct	Proportion of herbaceous wetland	
owpct	Proportion of open water	
hpctpreww, hpctpostww, hpct7089, hpctpost90	Percentage of houses built before WWII, after WWII, between 1970-1989, and after 1990	
whitepct, blackpct, asianpct, hispanicpct	Percentage of white, african american, asian, and hispanic population	

tot_pop	Total population	
Income	Median household income	

110

111 We performed the two-sample Kolmogorov-Smirnov (KS) test [30] as a univariate analysis to identify the
112 candidate key predictors to get some insights that would be helpful before building the models. The KS
113 test is a model-free approach to test whether the distributions of features corresponding to two different
114 classes behave similarly. Therefore, we did not reject collinearity in the KS test. A list of p-values was
115 calculated to assess the importance of the features. We examined the distributions of the weather,
116 socioeconomic, and land cover factors separately for the presence and absence of human cases by
117 hexagon and week. The KS test would indicate which variables are distributed differently for the two
118 situations. The KS score will serve as a criteria in choosing which variables to keep among a set of
119 highly-correlated variables. For each set of highly-correlated variables, we will keep the one with the
120 highest KS score.

121 Before building the light GBM model, we first assessed collinearity by generating the covariance plots
122 calculated from Pearson's correlation. We set the correlation threshold at 0.35 and kept the variables with
123 the largest $-\log(p)$ values in the KS test. Therefore, it is possible that the factors selected in the model are
124 correlated with the true predictors. We then selected the variables with the highest functional significance
125 to build the models. We also evaluated the income-stratified data to check whether the high-income and
126 low-income groups have different characteristics (S3).

127

128 Two-step Light GBM Modeling

129 The hyperparameter for the light GBM is tuned with grid search with a predefined set, evaluated on the
130 metric log-loss score as the decision criterion, which can help deal with the highly zero-inflated
131 characteristic of the WNV case number. We used the lightgbm package in Python [34] to perform the
132 light GBM method.

133 The model was built using a heuristic approach with two light GBM categorization procedures. After
134 removing the correlation, we ran the first light GBM procedure on all remaining variables. We then
135 examined the distribution of feature importance, selected the top variables by the natural gap in the
136 distribution, and ran another light GBM procedure. Feature importance is defined as the mean decrease in
137 impurity when a given feature is included to split the $WNV_binary = 0$ and $WNV_binary = 1$ cases.
138 Feature importance is represented by the negative logarithm of the absolute value of importance. We
139 evaluated the receiver operating characteristics area under the curve (ROC-AUC) to find the best
140 threshold for a minimum model. The ROC-AUC score is insensitive to imbalanced data. With the
141 threshold identified, we are able to evaluate the accuracy, recall, precision, and F-1 score [35]. We first fit
142 the model with high and low income data to confirm that the models are similar (S3). Therefore, we build
143 our model based on the full dataset. We then examine the distribution of feature importance and select
144 subsets of features to build reduced models. We examine the performance of the reduced models to find a
145 minimal model that retains predictive power.

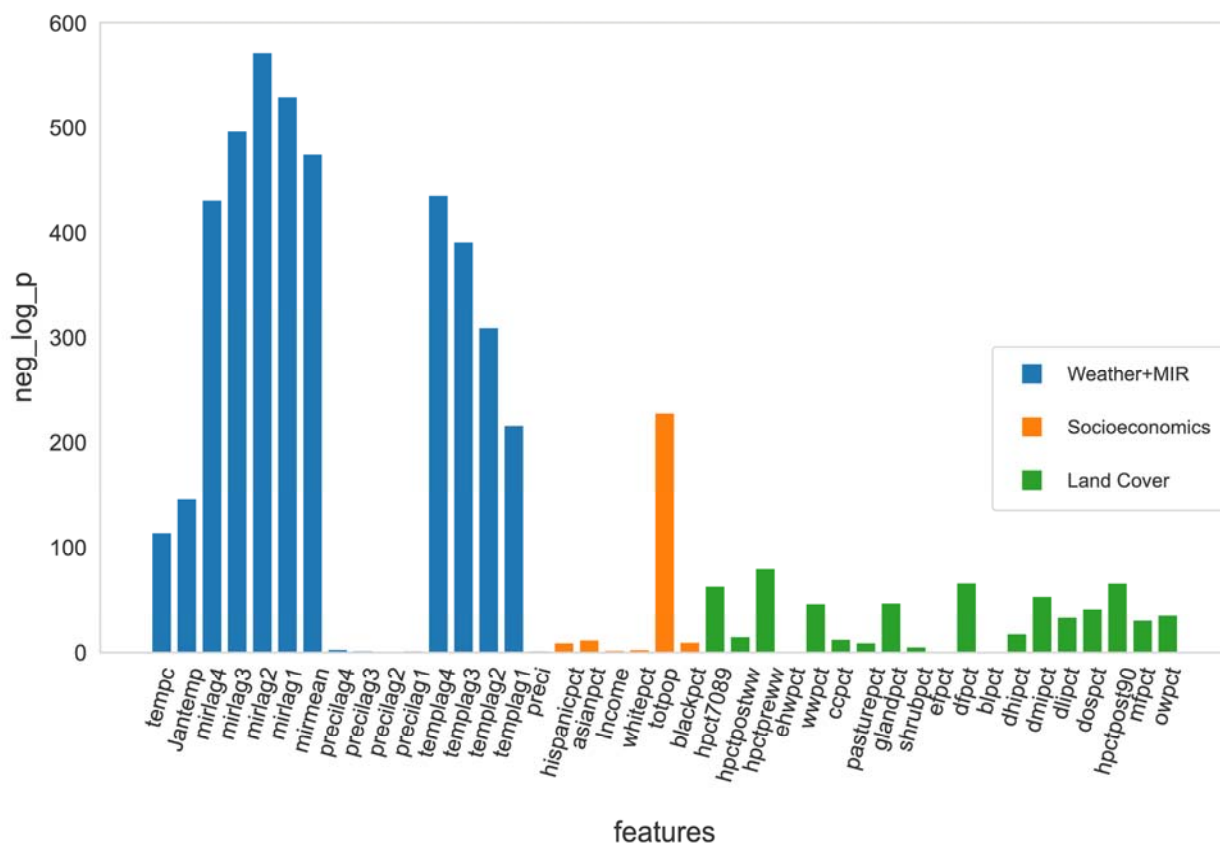
146 Then, in the final model, we evaluated the relative importance of the covariates to identify important
147 predictive features for WNV cases in our models. For the features of interest, we generate partial
148 dependence (PD) plots to show their marginal predicted probability. The slope of the PD plot represents
149 the strength of the feature. The shape of the PD plot could also indicate whether the effect is monotonic.
150 The PD plots could easily show the nonlinear effects that are difficult to identify by regression.

151

152 Results

153 KS test

154 We performed univariable KS tests on all variables (Figure 1). We found that temperatures and mosquito
 155 infection rates have significant effects in the model. On the other hand, precipitation, land cover and
 156 socio-economic characteristics do not contribute significantly to the WNV risk.



157

158 **Fig 1. $-\log(p)$ of Kolmogorov-Smirnov test for all the features and covariates.** From the KS test, we
 159 calculate the p-value, which indicates how different the distribution of the variable is between the
 160 hexagon-weeks with and without a case. The larger the $-\log(p)$, the less similar the two distributions are.
 161 The variables are grouped into four main categories + one residue category, but we have combined the
 162 fast-changing weather and MIR into the same category because these variables, as well as the number of

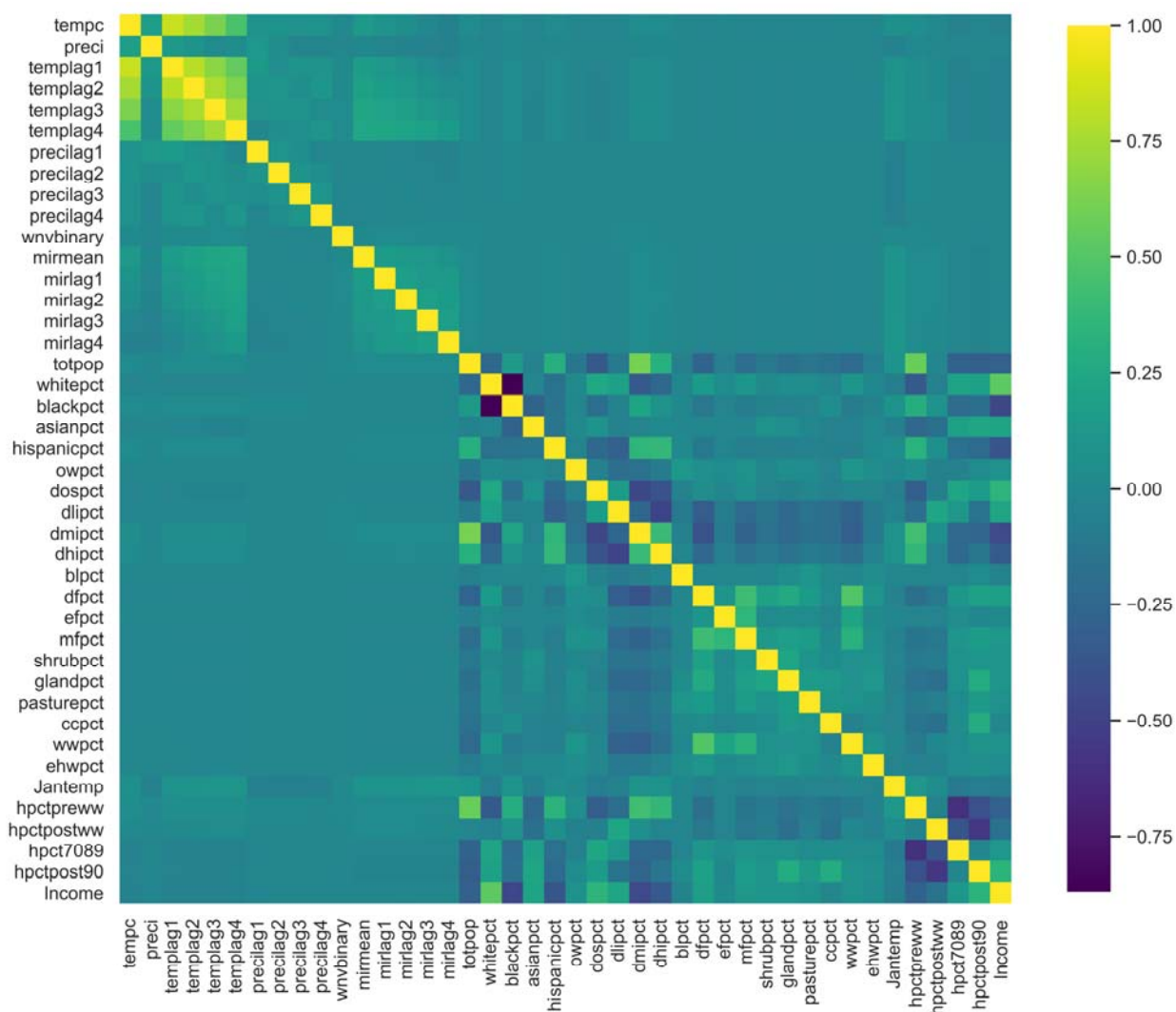
163 cases, are captured with a temporal resolution of one week. Green bars represent land cover variables.

164 Orange bars represent socio-economic variables. The blue bars represent strongly fluctuating variables:

165 weather and mosquito infection rate.

166

167 Variable correlations



169 **Fig 2. Heat-Map Covariance matrix for all the features.** Original data are from Karki (2020) [26]. Yellow

170 colors indicate strong positive correlations; dark blue colors indicate strong negative correlations. Light blue or

171 green colors indicate weak correlations. We infer that temperature has a relatively high temporal correlation, as
172 the variables tempc and temp1-4 (current temperature and temperatures 1-4 weeks before) are correlated. In
173 addition, development stage and housing age are correlated with population, showing the interaction of
174 population aggregation with land cover and housing status.

175 Figure 2 shows the correlation between the variables. We found that weekly temperatures have a strong
176 positive temporal correlation (0.47 - 0.84). On the other hand, the lagged effects of weekly MIR (0.075 -
177 0.18) and weekly precipitation (-0.022 - 0.044) are not as strongly correlated. Weekly MIR and weekly
178 precipitation are also independent of other variables.

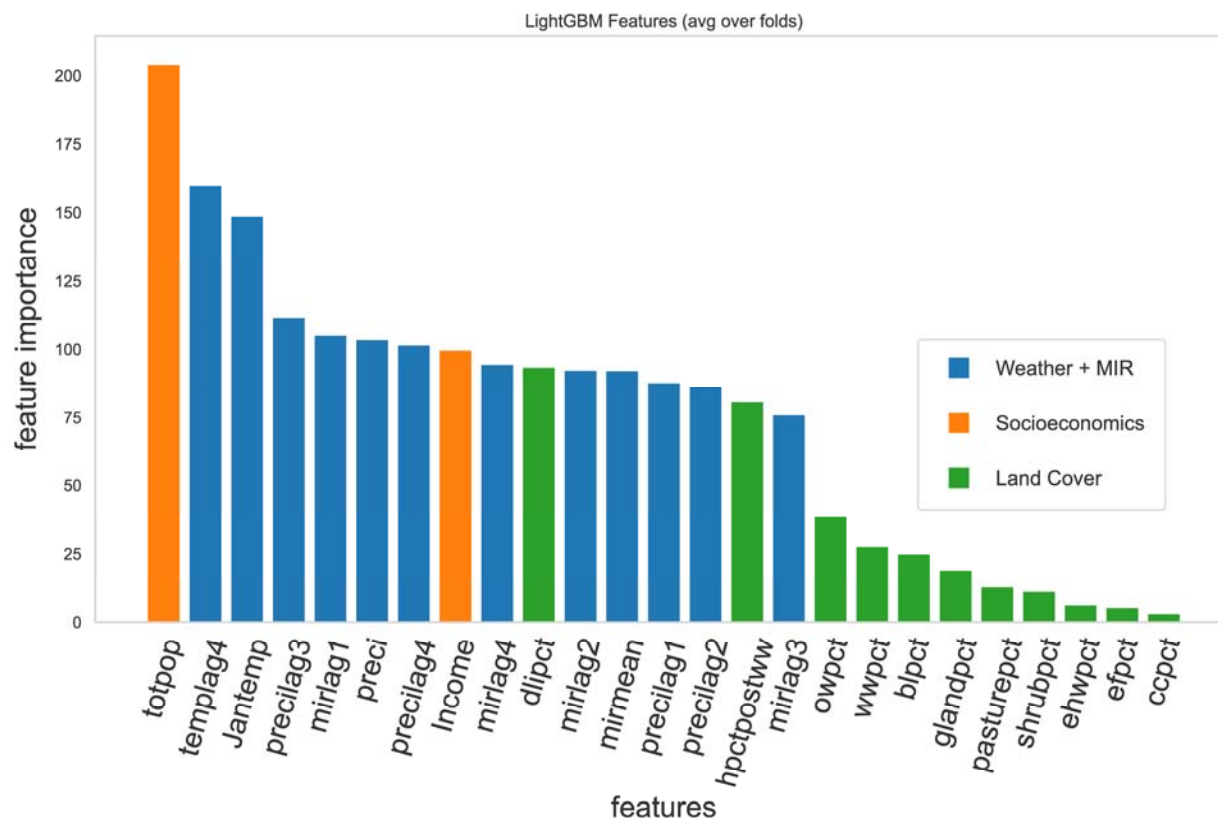
179 We also found that income is strongly correlated with race. Income has a high positive correlation (0.54)
180 with the white race percentage in the hexagon area, and a medium-high negative correlation with the
181 black race percentage (-0.46) and the Hispanic race percentage (-0.37). The white and black population
182 percentages have a strong negative correlation with each other (-0.87), which is to be expected since the
183 total population percentages should add up to 100%.

184 For each set of medium to highly correlated variables, we kept the variables with the highest KS scores
185 for the light GBM analysis. The remaining variables are: All precipitation and MIR variables, mean
186 temperature of 4 weeks before the human case report, mean temperature in January, total population,
187 proportion of developed low intensity, proportion of open water, proportion of barren land, proportion of
188 evergreen forest, proportion of shrubs, proportion of grassland, proportion of pasture, proportion of
189 cultivated land, proportion of woody wetlands, emergent herbaceous wetlands, percent temperature in
190 January, house post World War II, and income.

191

192 Light GBM based on all selected features

193 We built our models using cross-validation, randomly splitting training and test sets, and then selected the
194 best parameter based on the log-loss criteria. The importance of each predictor in the model is shown in
195 Figure 3, and its performance on the test set is shown in Table 2.



196
197 **Fig 3: Gini feature importance of the model predicting West Nile Virus cases in the Chicago area,**
198 **with the 25 variables after removing the highly correlated ones.** The higher the y-value, the more
199 important the feature is to the model. The variables are grouped into four main categories, but we have
200 combined the fast-changing weather and MIR into the same category because these variables, as well as
201 the number of cases, are captured with a temporal resolution of one week. The blue bars represent the
202 weekly variables (weather + MIR). Orange bars represent socio-economic variables. The green bars

203 represent the land cover variables. We found that total population is the most important variable in the
204 model. The weekly variables (weather + MIR) are also strong predictors.

205 Figure 3 shows that socioeconomic, weather, and mosquito infection factors are candidates for strong
206 predictors. Precipitation variables have relatively low importance among the weather factors, but still
207 have a medium rank among the feature importance. Total population and income level, the two
208 independent socioeconomic variables included in the model, both have high importance in predicting
209 WNV case occurrence. Percentage of housing built after World War II and percentage of low
210 development intensity area are the only strong indicators among the land cover features. We can see a
211 natural gap between MIR 3 weeks before (mirlag3) and percentage of open water (owpct). Therefore, we
212 select the first 16 features for our reduced model.

213 The cutoff for selecting the features is chosen to maximize the difference between the true positive rate
214 (TPR) and the false positive rate (FPR). Table 2 shows the confusion matrix of the result based on the test
215 set. With the cutoff = 0.625, we obtain a true positive rate (recall or sensitivity) close to 0.89. The
216 precision is about 0.007. This value is not good, but it is still well above the baseline derived from the
217 proportion of positive categories (0.0005) in the dataset. The F1 score is 0.486 and the accuracy is 0.92.
218 Since our model focuses on maximizing recall, this loss in overall performance is to be expected.

219 **Table 2: Confusion Matrix of the model including all features.**

	hexagons with WNV cases predicted	hexagons with no WNV cases predicted
Hexagons with WNV Case Observed	160	18
Hexagons with no WNV Case	22878	265923

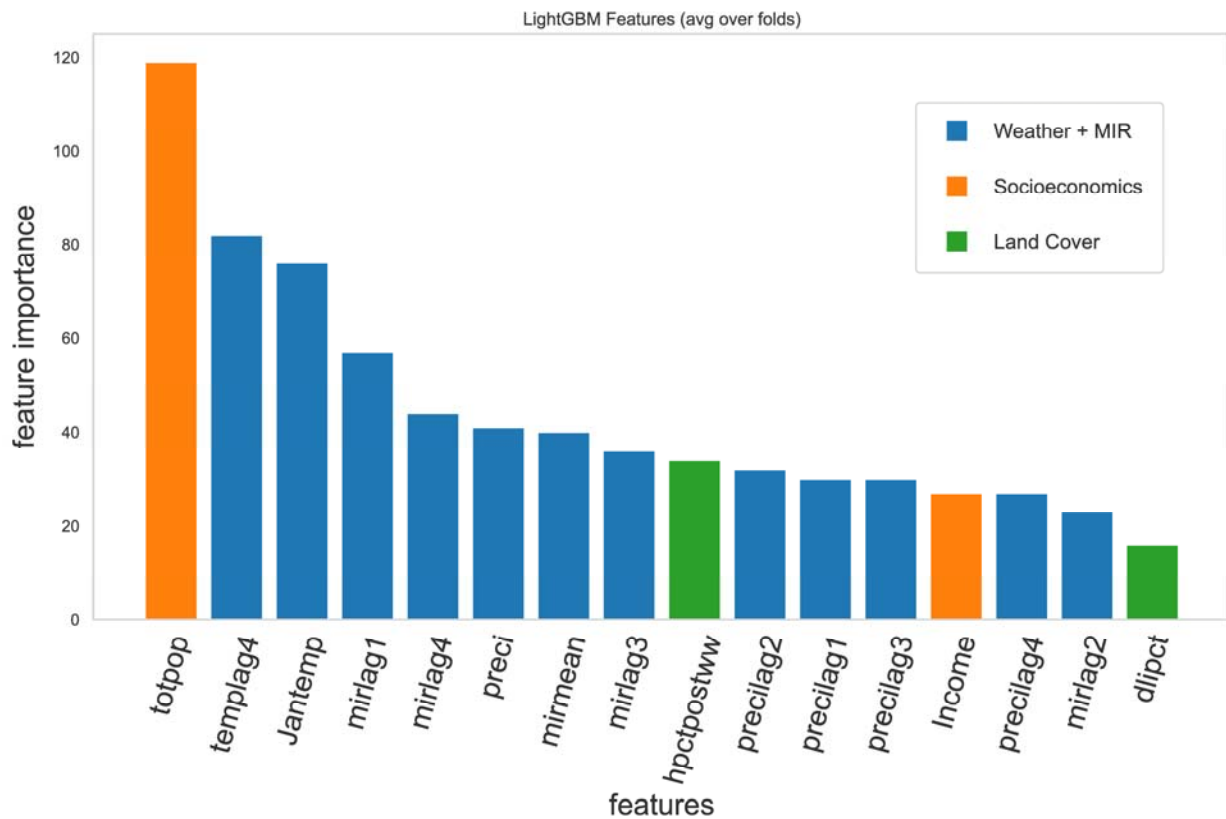
Observed		
----------	--	--

220 We predict the probability that a case of WNV will occur during a given week in each 1-km-wide
221 hexagonal region in Cook and DuPage counties, from which we predict whether a case will occur. The
222 receiver operating characteristic (ROC) area under the curve (AUC) is 0.96. The model has an accuracy
223 of 0.92, a precision of 0.007, a recall of 0.89, and a macro F1 score of 0.486.

224

225 Light GBM model based on reduced features

226 We re-fit the model using only the features with importance > 20. The feature importance of each
227 predictor in this model is shown in Figure 4, and its performance on the test set is shown in Table 2.



228

229 **Fig 4: Gini Feature importance of the candidate predictors in the reduced model.** Blue, highly dynamic
230 features including weather and mosquito infection rate. Orange: Socioeconomic features. Green, land cover
231 data. The socioeconomic features include total population and income, ranked 1 and 5. The land cover
232 features, share of post-war housing and share of low-intensity development, rank 11 and 15. The most
233 important natural features are the January temperature and the average weekly temperature, followed by the
234 mosquito infection rate. While the ranks may change in individual runs, the feature importance of these factors
235 are close to each other.

236 The cutoff for selecting the features is chosen to maximize the difference between the true positive rate
237 (TPR) and the false positive rate (FPR). Table 3 shows the confusion matrix of the result based on the test
238 set. With the cutoff = 0.446, we obtain a true positive rate (recall or sensitivity) close to 0.96. The
239 precision is about 0.0034. This value is not good, but it is still well above the baseline derived from the
240 proportion of positive categories (0.0005) in the dataset. The F1 score is 0.45 and the accuracy is 0.83.
241 Since our model focuses on maximizing recall, this loss in overall performance is to be expected.

242 **Table 3: Confusion Matrix of the reduced model.**

	hexagons with WNV cases predicted	hexagons with no WNV cases predicted
Hexagons with WNV Case Observed	173	6
Hexagons with no WNV Case Observed	50,060	238,741

243
244 We predict the probability that a case of WNV will occur during a given week in each 1-km-wide
245 hexagonal region in Cook and DuPage counties, from which we predict whether a case will occur. The

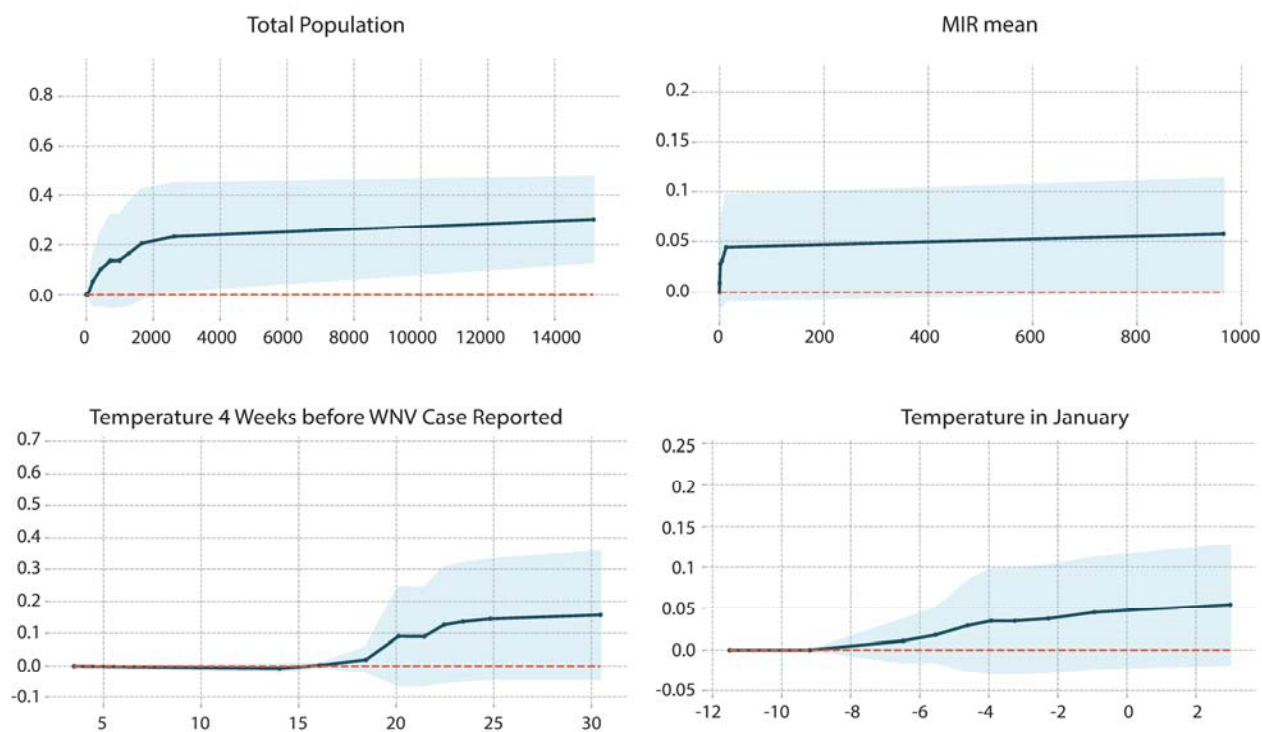
246 receiver operating characteristic (ROC) area under the curve (AUC) is 0.95. The model has an accuracy
247 of 0.8267, a precision of 0.0034, a recall of 0.9664, and a macro F1 score of 0.45.

248 Based on the above results, we found that the metrics of the reduced model are similar to the model
249 including all 25 low-correlation variables. Therefore, we conclude that the reduced model is sufficient to
250 describe the result.

251

252 Marginal Effects

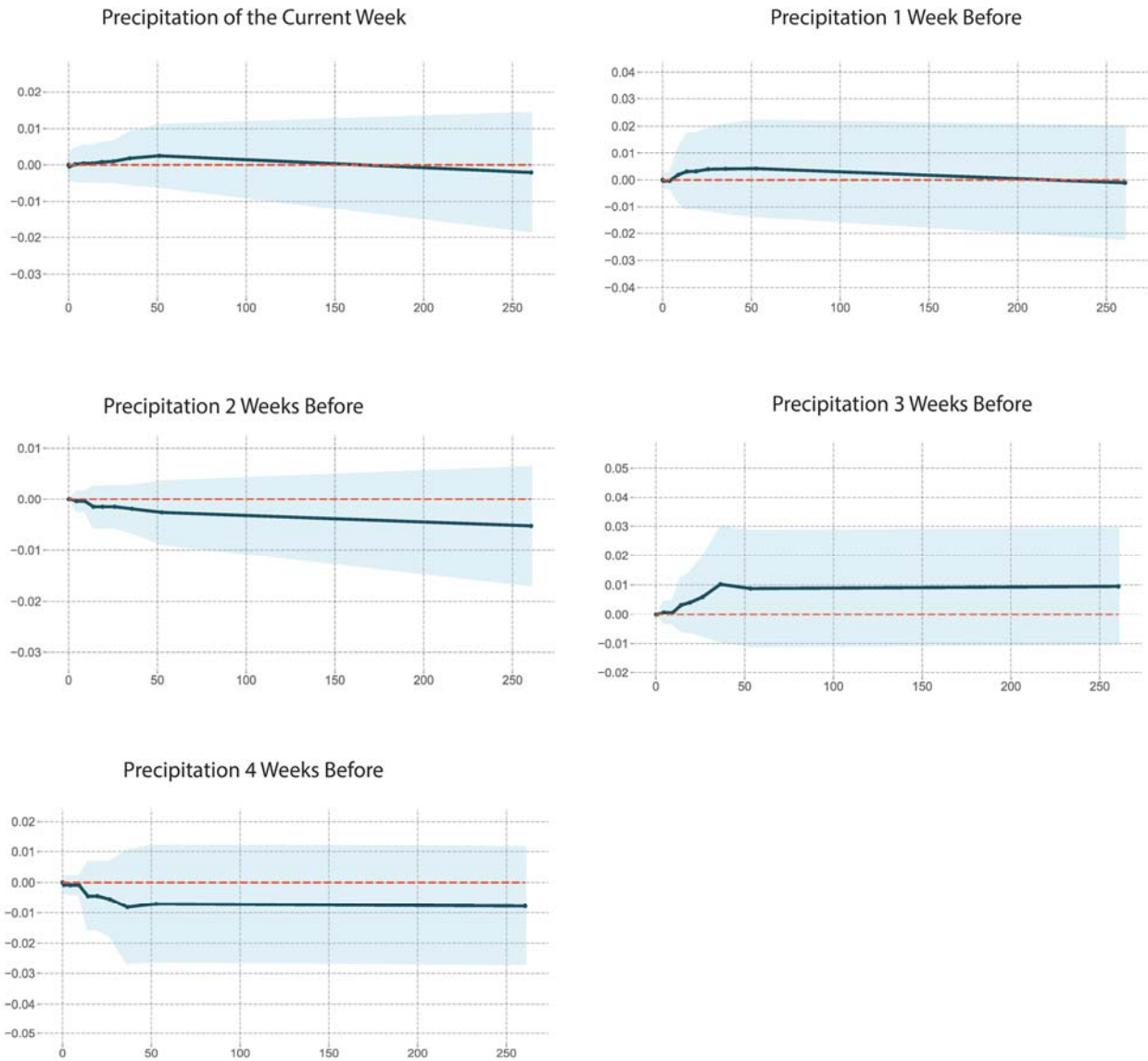
253 We examined the marginal effects of all the features by generating partial dependence plots. The slope of
254 the plots shows how much each feature contributes to the model when controlling for the other factors.



255
256 **Fig 5. Partial dependence plot of factors with positive effects: total population, mean MIR,**
257 **temperature 4 weeks before WNV cases are reported, and January temperature. The MIR and the**

258 weekly temperatures in 1-4 weeks before also have similar trends as the mean MIR and the temperature
259 of the current week.

260 Figure 5 shows the partial dependence plot of the factors that predict higher WNV risk as the values of
261 the factors increase. MIR and total population have strong monotonic positive effects. The result is
262 consistent that both disease-carrying mosquitoes and the human population increase the risk of infection.
263 Weekly mean temperature 4 weeks before WNV cases are reported has a strong monotonic positive
264 effect. It is noteworthy that the temperature range is below 30°C, which is approximately the range that
265 promotes mosquito activity and virus replication. January temperature also has a strong monotonic
266 positive effect. A warmer January allows mosquitoes to survive the winter, resulting in larger mosquito
267 populations.



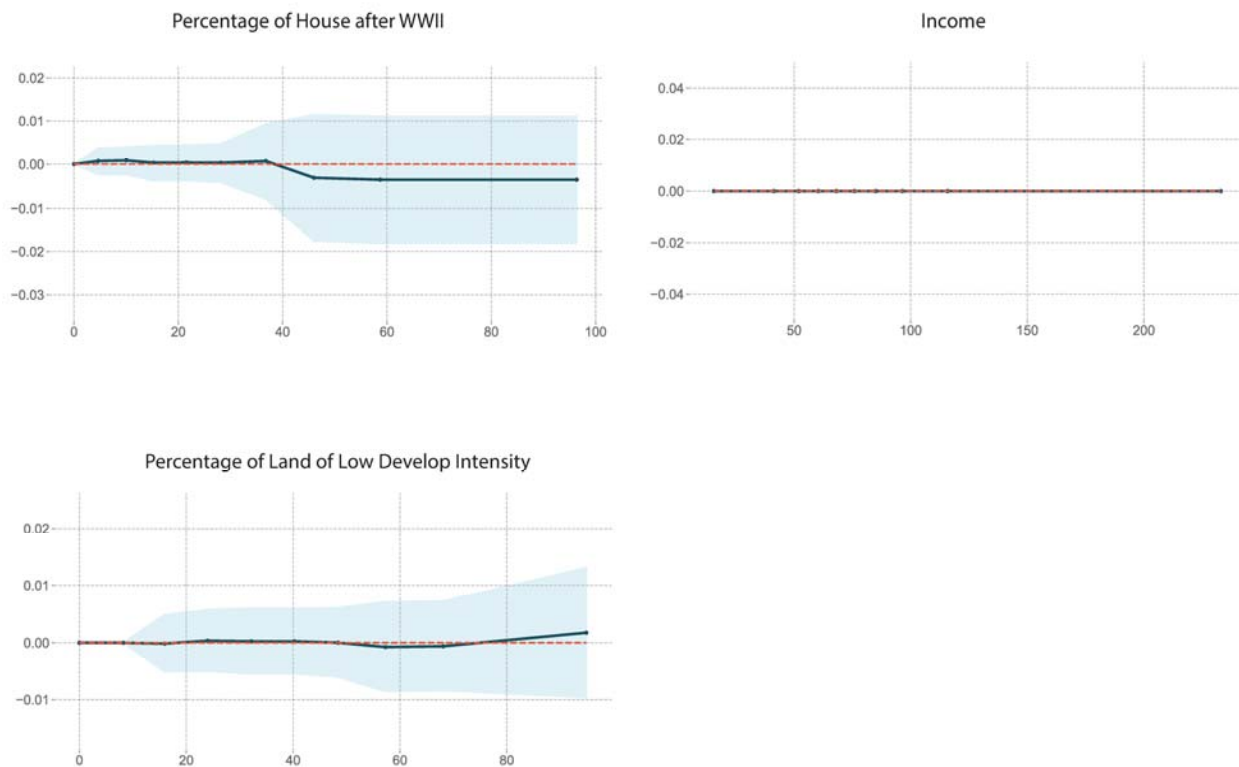
268

269 **Fig 6. Partial dependence plot of precipitation for the current week and 1-4 weeks prior.**

270 Precipitation variables have non-monotonic effects. The marginal effect contributing to WNV risk first
271 increases and then decreases as precipitation increases.

272 On the other hand, as shown in Figure 6, the precipitation variables have non-monotonic effects, i.e., the
273 risk of WNV outbreak first increases and then decreases as precipitation continues to increase. This result
274 is consistent with the existing literature [10,14,26]. While temporary water accumulation provides

275 mosquitoes with more places to lay eggs, excessive precipitation can also wash away mosquito eggs, thus
276 reducing the risk of WNV.



277
278 **Fig 7. Partial dependence plot of socioeconomic and land cover features.** The socioeconomic and
279 land cover features are not very strongly represented. There is no marginal effect of income. The
280 percentage of houses built after World War II has a slight negative effect, indicating that people living in
281 older neighborhoods have higher WNV risks. Meanwhile, the percentage of less developed land has a
282 slight positive effect.

283 As shown in Figure 7, apart from total population (in Figure 1), land cover and other socioeconomic features
284 have relatively small effects. We don't observe a marginal effect of income, although it is presented in the
285 feature selection. House age and land development intensity both have small effects on WNV case prediction.

286

287 Conclusion

288 We performed two-step light GBM procedures to identify a minimum model. We evaluated the ROC-
289 AUC score, accuracy, recall, precision and F-1 score of the models. We found that the reduced model has
290 a worse performance than the linear models of Karki, et. al. [26], while the full model has a similar
291 performance. Therefore, we kept all 25 parameters in the model for prediction. We have found that the
292 natural effects including January temperature, weekly temperature (lagged 0-4 weeks), weekly
293 precipitation (lagged 0-4 weeks), and weekly MIR (lagged 0-4 weeks), as well as the total population are
294 the dominant features that are strongly correlated with the incidence of West Nile virus human cases.

295 The light GBM model is better at detecting the positive cases, i.e. higher recall. We found consistent
296 features with Karki, et al. that mosquito infection rate, temperature and their lag effects are important
297 factors [26]. This result was further confirmed with PD plots. We also found the behavior of precipitation
298 factors consistent with the literature [10,14,26], being strong predictors with non-monotonic marginal
299 effects. In addition, we found that the percentage of houses built after World War II, which is not
300 included in the original work, is quite important. While income is selected as a predictor by the final
301 model, the PD plot has shown that it has no marginal effects.

302 The model based only on selected key factors performs similarly to the model that includes all other
303 factors. In addition, both the number of cases and the weather vary on a weekly basis, while the land
304 cover and socioeconomic data are static. Therefore, the effect of the socioeconomic characteristics could
305 be masked by the correlated characteristics of lagged MIR and lagged temperature.

306 One concern was that the behavior of the model may differ by the income of the area, as income
307 disparities may affect diagnosis rates, surveillance efforts, and distribution of land cover and housing
308 variables. Therefore, the light GBM model fitting was repeated for subsets of the data consisting of the
309 areas with above-median income and the areas with below-median income (S3). These stratified models

310 were similar to each other and to the full model, indicating that the predictive capabilities of this model
311 are not predicated on income groupings.

312 In conclusion, our light GBM model provides an alternative way to predict the probability of an area
313 having a WNV case or not. The performance in terms of ROC-AUC is very close to the previous work
314 [26] and is much better at detecting the area where there is actually a case. We also have a clearer
315 relationship between temperature and precipitation, mosquito infection, and West Nile virus. In addition,
316 we identified weak effects of socioeconomics and land cover. The risk of contracting WNV does not
317 appear to be related to income in these data. However, other factors may relate to income and WNV
318 detection that are not possible to study with these data, such as variation in diagnosis rates.

319 The results of this study can be used as a guideline to develop a threshold for public health intervention.

320 Acknowledgements

321 The authors thank the NCSA Center-Directed Discretionary Research (CDDR) for funding this project. The
322 authors would like to thank the SPIN program at NCSA for supporting the student who is the first author of the
323 paper. The authors would like to thank the HAL cluster and support team for providing the computational
324 resources to complete the work. The author would also like to acknowledge the efforts of the NCSA Industry
325 Group for supporting the work. The authors would like to thank Dr. Christina Fliege for her editorial
326 suggestions on this manuscript. The authors would like to thank Mr. Mingyu Yang for his help in retrieving
327 and preprocessing the census data.

328

329 References

330

331 1. Lanciotti RS. Origin of the West Nile Virus Responsible for an Outbreak of Encephalitis in the
332 Northeastern United States. *Science*. 1999. pp. 2333–2337. doi:10.1126/science.286.5448.2333

333 2. Hayes EB, Komar N, Nasci RS, Montgomery SP, O’Leary DR, Campbell GL. Epidemiology and
334 transmission dynamics of West Nile virus disease. *Emerg Infect Dis*. 2005;11: 1167–1173.

335 3. Hadfield J, Brito AF, Swetnam DM, Vogels CBF, Tokarz RE, Andersen KG, et al. Twenty years of
336 West Nile virus spread and evolution in the Americas visualized by Nextstrain. *PLoS Pathog*.
337 2019;15: e1008042.

338 4. Kilpatrick AM, Marm Kilpatrick A, LaDeau SL, Marra PP. ECOLOGY OF WEST NILE VIRUS
339 TRANSMISSION AND ITS IMPACT ON BIRDS IN THE WESTERN HEMISPHERE. *The Auk*.
340 2007. p. 1121. doi:10.1642/0004-8038(2007)124[1121:eownvt]2.0.co;2

341 5. Kramer LD, Styer LM, Ebel GD. A Global Perspective on the Epidemiology of West Nile Virus.
342 *Annual Review of Entomology*. 2008. pp. 61–81. doi:10.1146/annurev.ento.53.103106.093258

343 6. Johnson BJ, Munafo K, Shappell L, Tsipoura N, Robson M, Ehrenfeld J, et al. The roles of mosquito
344 and bird communities on the prevalence of West Nile virus in urban wetland and residential habitats.
345 *Urban Ecosystems*. 2012. pp. 513–531. doi:10.1007/s11252-012-0248-1

346 7. Reisen WK. Ecology of West Nile virus in North America. *Viruses*. 2013;5: 2079–2105.

347 8. Hubálek Z, Halouzka J. West Nile Fever—a Reemerging Mosquito-Borne Viral Disease in Europe.
348 *Emerging Infectious Diseases*. 1999. pp. 643–650. doi:10.3201/eid0505.990505

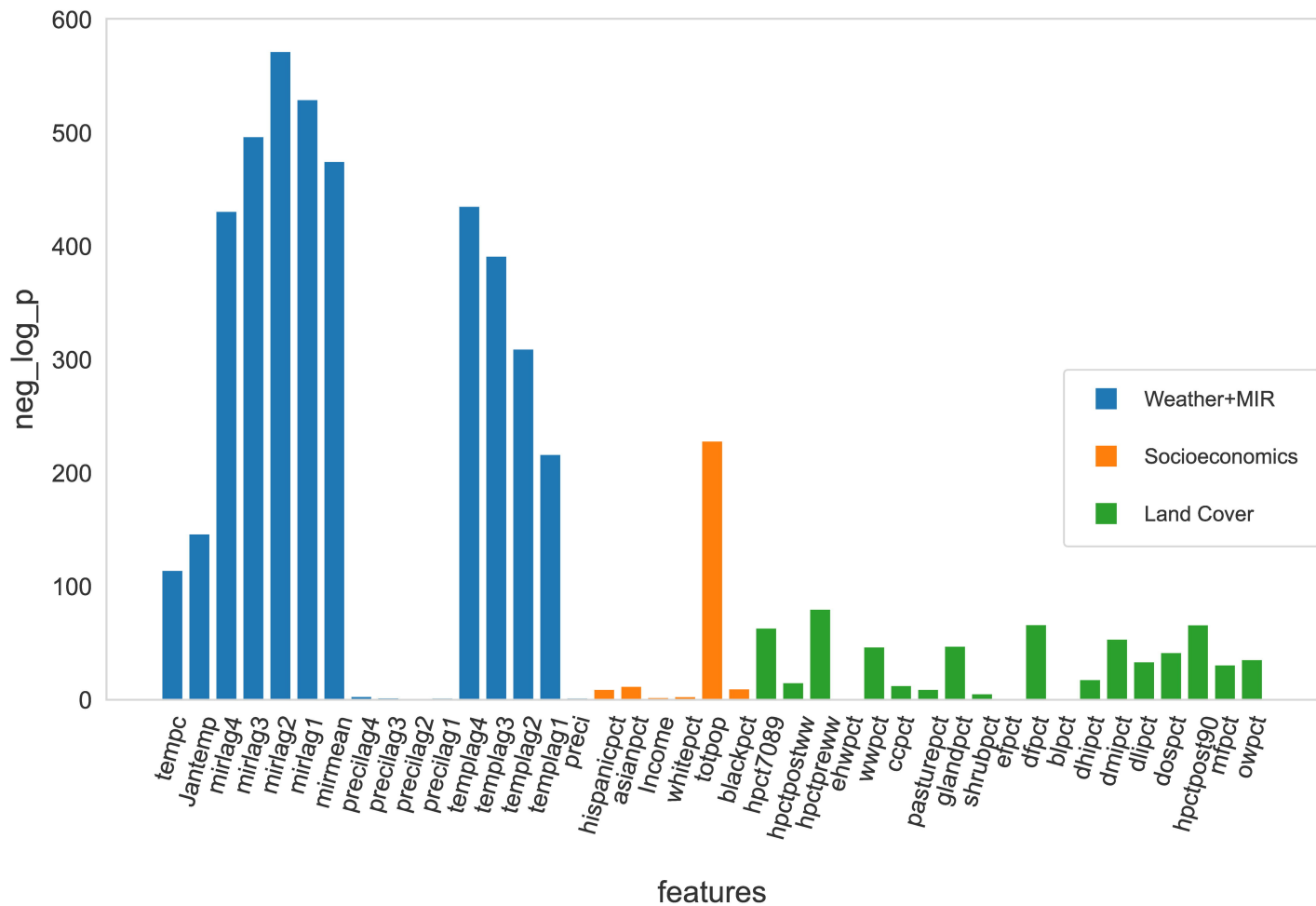
349 9. Kilpatrick AM, Pape WJ. Predicting human West Nile virus infections with mosquito surveillance

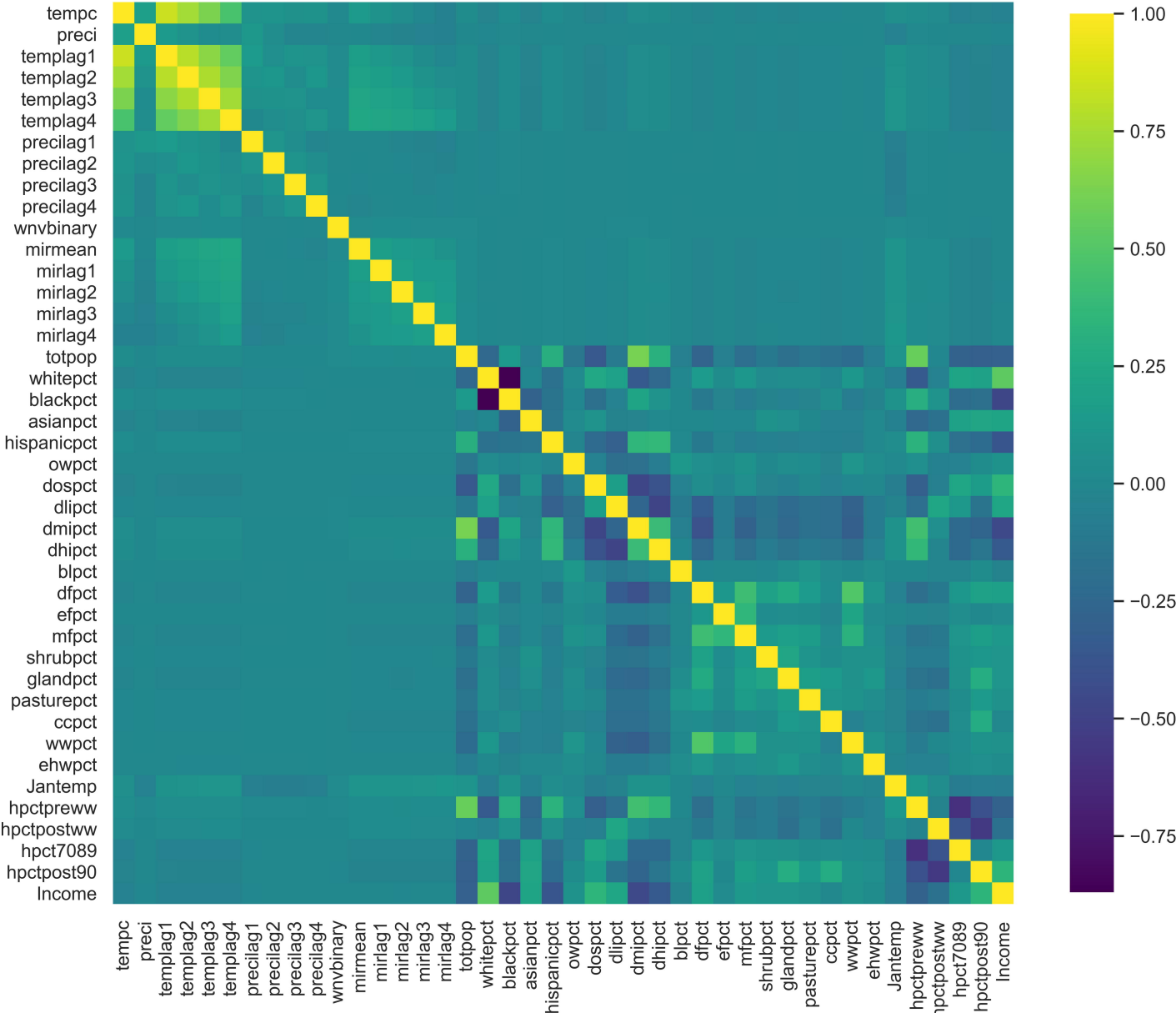
- 350 data. *Am J Epidemiol.* 2013;178: 829–835.
- 351 10. Keyel AC, Elison Timm O, Backenson PB, Prussing C, Quinones S, McDonough KA, et al.
352 Seasonal temperatures and hydrological conditions improve the prediction of West Nile virus
353 infection rates in *Culex* mosquitoes and human case counts in New York and Connecticut. *PLoS*
354 *One.* 2019;14: e0217854.
- 355 11. Paz S. Effects of climate change on vector-borne diseases: an updated focus on West Nile virus in
356 humans. *Emerging Topics in Life Sciences.* 2019. pp. 143–152. doi:10.1042/etls20180124
- 357 12. Hahn MB, Nasci RS, Delorey MJ, Eisen RJ, Monaghan AJ, Fischer M, et al. Meteorological
358 Conditions Associated with Increased Incidence of West Nile Virus Disease in the United States,
359 2004–2012. *The American Journal of Tropical Medicine and Hygiene.* 2015. pp. 1013–1022.
360 doi:10.4269/ajtmh.14-0737
- 361 13. Shocket MS, Verwillow AB, Numazu MG, Slamani H, Cohen JM, El Moustaid F, et al.
362 Transmission of West Nile and five other temperate mosquito-borne viruses peaks at temperatures
363 between 23°C and 26°C. *eLife.* 2020. doi:10.7554/elife.58511
- 364 14. Shand L, Brown WM, Chaves LF, Goldberg TL, Hamer GL, Haramis L, et al. Predicting West Nile
365 Virus Infection Risk From the Synergistic Effects of Rainfall and Temperature. *J Med Entomol.*
366 2016;53: 935–944.
- 367 15. Poh KC, Chaves LF, Reyna-Nava M, Roberts CM, Fredregill C, Bueno R, et al. The influence of
368 weather and weather variability on mosquito abundance and infection with West Nile virus in Harris
369 County, Texas, USA. *Science of The Total Environment.* 2019. pp. 260–272.
370 doi:10.1016/j.scitotenv.2019.04.109
- 371 16. Champion M, Bina C, Pozniak M, Hanson T, Vaughan J, Mehus J, et al. Predicting West Nile Virus
372 (WNV) occurrences in North Dakota using data mining techniques. 2016 *Future Technologies*

- 373 Conference (FTC). 2016. doi:10.1109/ftc.2016.7821628
- 374 17. Peper ST, Dawson DE, Dacko N, Athanasiou K, Hunter J, Loko F, et al. Predictive Modeling for
375 West Nile Virus and Mosquito Surveillance in Lubbock, Texas. *J Am Mosq Control Assoc.* 2018;34:
376 18–24.
- 377 18. Davis JK, Vincent GP, Hildreth MB, Kightlinger L, Carlson C, Wimberly MC. Improving the
378 prediction of arbovirus outbreaks: A comparison of climate-driven models for West Nile virus in an
379 endemic region of the United States. *Acta Tropica.* 2018. pp. 242–250.
380 doi:10.1016/j.actatropica.2018.04.028
- 381 19. Yoo E-H, Chen D, Diao C, Russell C. The Effects of Weather and Environmental Factors on West
382 Nile Virus Mosquito Abundance in Greater Toronto Area. *Earth Interactions.* 2016. pp. 1–22.
383 doi:10.1175/ei-d-15-0003.1
- 384 20. DeFelice NB, Birger R, DeFelice N, Gagner A, Campbell SR, Romano C, et al. Modeling and
385 Surveillance of Reporting Delays of Mosquitoes and Humans Infected With West Nile Virus and
386 Associations With Accuracy of West Nile Virus Forecasts. *JAMA Netw Open.* 2019;2: e193175.
- 387 21. Sánchez-Gómez A, Amela C, Fernández-Carrión E, Martínez-Avilés M, Sánchez-Vizcaíno JM,
388 Sierra-Moros MJ. Risk mapping of West Nile virus circulation in Spain, 2015. *Acta Trop.* 2017;169:
389 163–169.
- 390 22. Hernandez E, Torres R, Joyce AL. Environmental and Sociological Factors Associated with the
391 Incidence of West Nile Virus Cases in the Northern San Joaquin Valley of California, 2011–2015.
392 *Vector-Borne and Zoonotic Diseases.* 2019. pp. 851–858. doi:10.1089/vbz.2019.2437
- 393 23. Myer MH, Johnston JM. Spatiotemporal Bayesian modeling of West Nile virus: Identifying risk of
394 infection in mosquitoes with local-scale predictors. *Sci Total Environ.* 2019;650: 2818–2829.

- 395 24. Farooq Z, Sjödin H, Semenza JC, Tozan Y, Sewe MO, Wallin J, et al. European projections of West
396 Nile virus transmission under climate change scenarios. *One Health*. 2023;16: 100509.
- 397 25. Bassal R, Shohat T, Kaufman Z, Mannasse B, Shinar E, Amichay D, et al. The seroprevalence of
398 West Nile Virus in Israel: A nationwide cross sectional study. *PLoS One*. 2017;12: e0179774.
- 399 26. Karki S, Brown WM, Uelmen J, Ruiz MO, Smith RL. The drivers of West Nile virus human illness
400 in the Chicago, Illinois, USA area: Fine scale dynamic effects of weather, mosquito infection, social,
401 and biological conditions. *PLoS One*. 2020;15: e0227160.
- 402 27. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient
403 Boosting Decision Tree. *Adv Neural Inf Process Syst*. 2017;30. Available:
404 <https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>
- 405 28. Breiman L. Random Forests. *Mach Learn*. 2001;45: 5–32.
- 406 29. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *aoas*. 2008;2: 841–
407 860.
- 408 30. Gong R & Huang. A Kolmogorov–Smirnov statistic based segmentation approach to learning from
409 imbalanced datasets: With application in property refinance prediction. *Expert Syst Appl*. 2012;39:
410 6192–6200.
- 411 31. Daly C, Smith JI, Olson KV. Mapping Atmospheric Moisture Climatologies across the
412 Conterminous United States. *PLoS One*. 2015;10: e0141140.
- 413 32. Dewitz J. National Land Cover Database (NLCD) 2019 Products. U.S. Geological Survey; 2021.
414 doi:10.5066/P9KZCM54
- 415 33. US Census Bureau. Census.gov. [cited 11 Aug 2020]. Available: <https://www.census.gov/en.html>

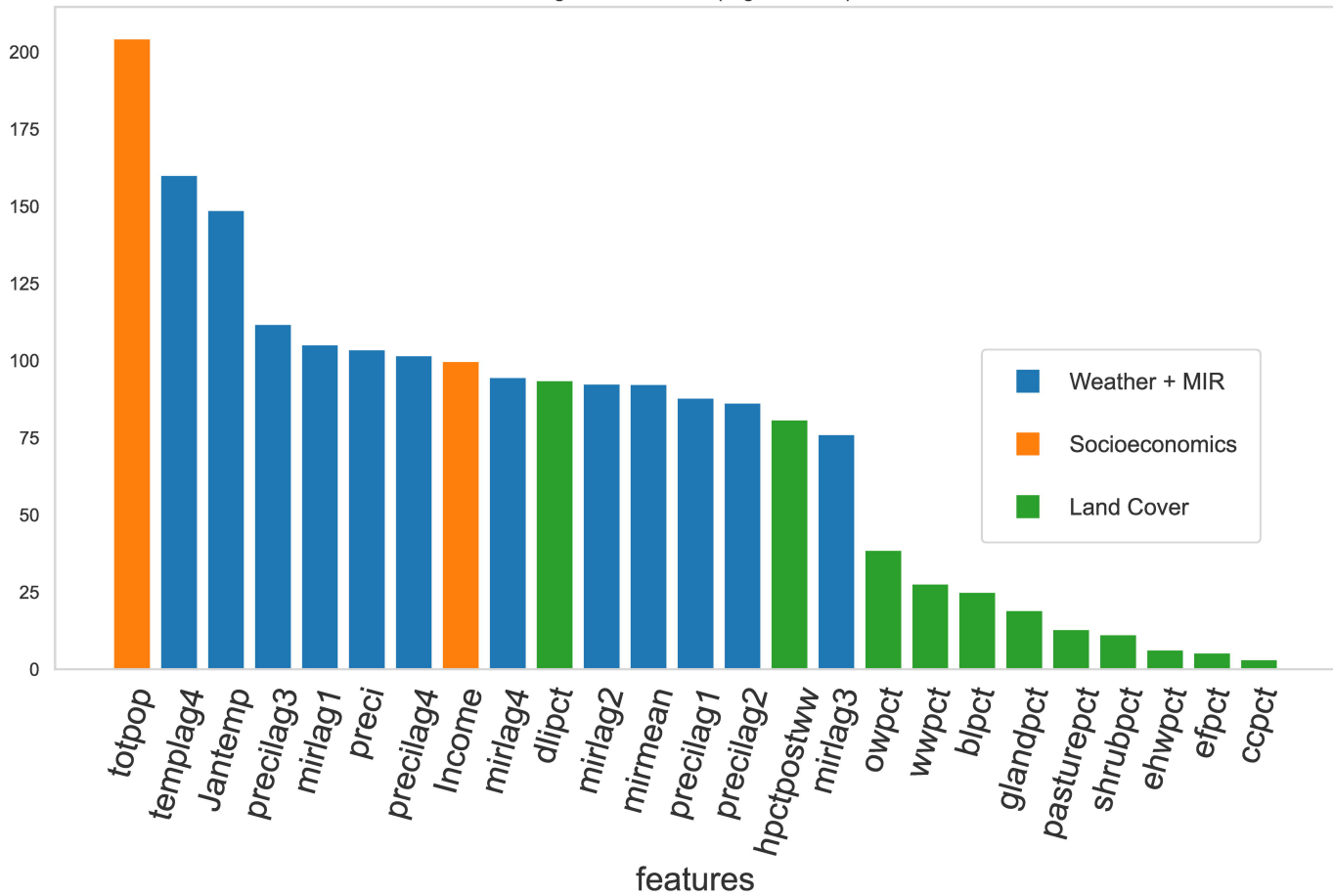
- 416 34. Machado MR, Karray S, de Sousa IT. LightGBM: an Effective Decision Tree Gradient Boosting
417 Method to Predict Customer Loyalty in the Finance Industry. 2019 14th International Conference on
418 Computer Science & Education (ICCSE). 2019. doi:10.1109/iccse.2019.8845529
- 419 35. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When
420 Evaluating Binary Classifiers on Imbalanced Datasets. PLoS One. 2015;10: e0118432.



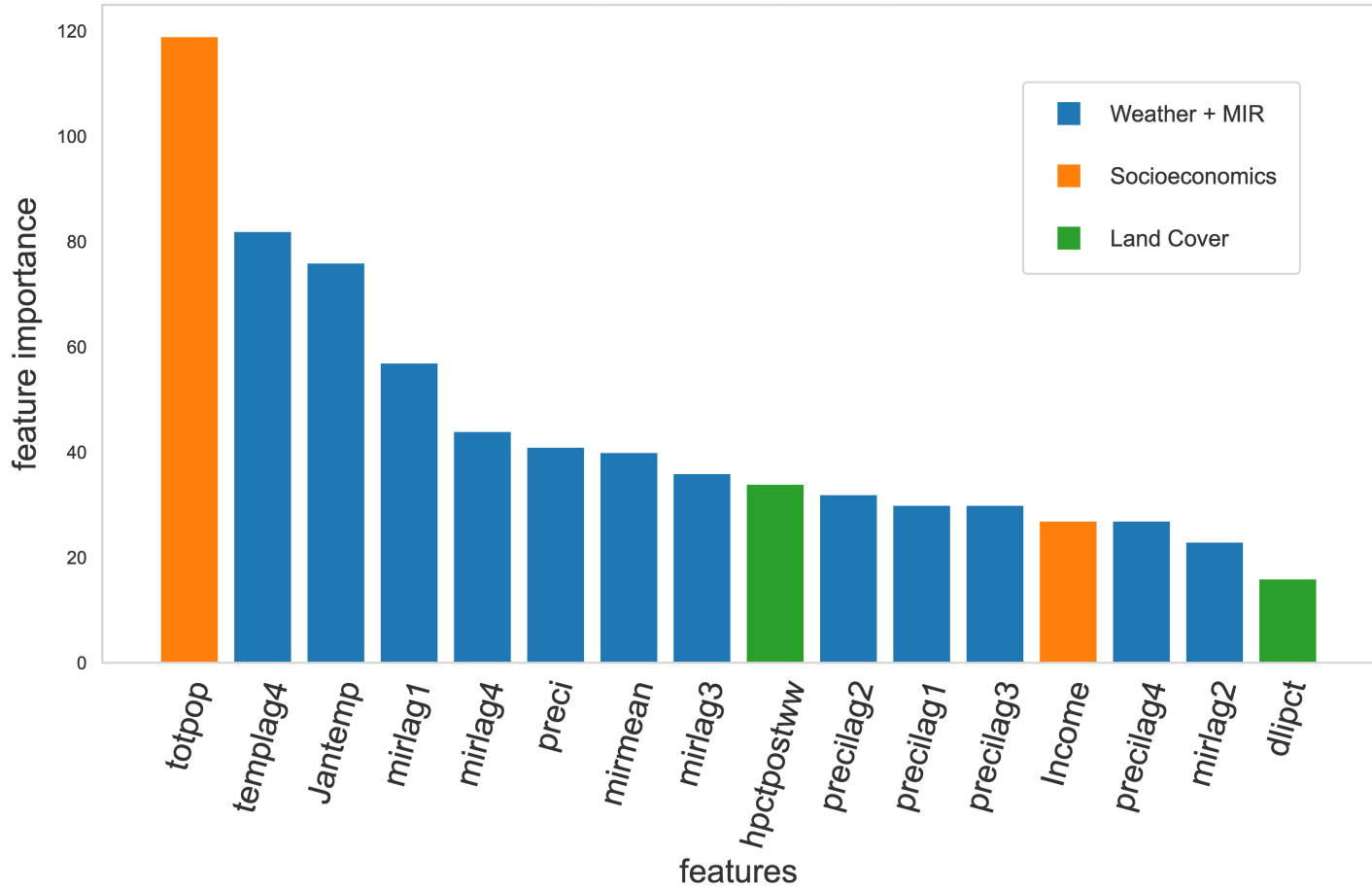


LightGBM Features (avg over folds)

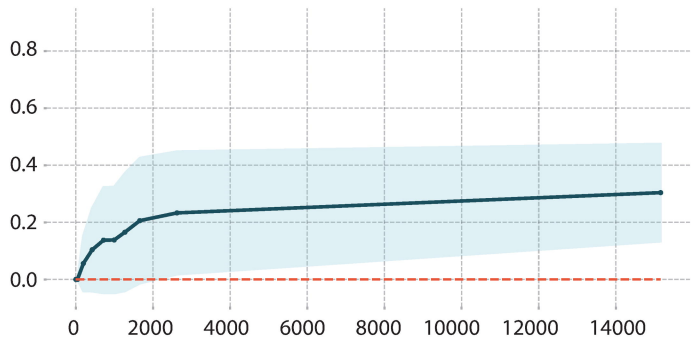
feature importance



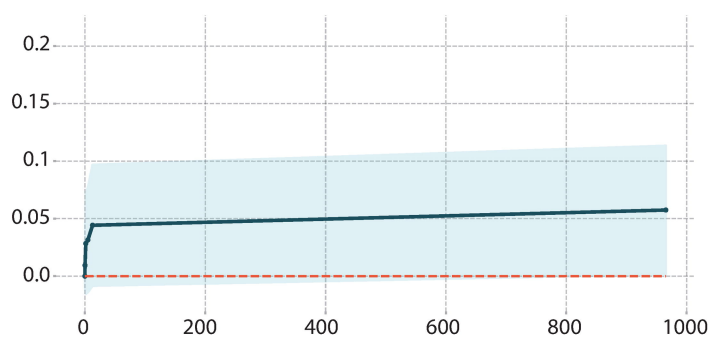
LightGBM Features (avg over folds)



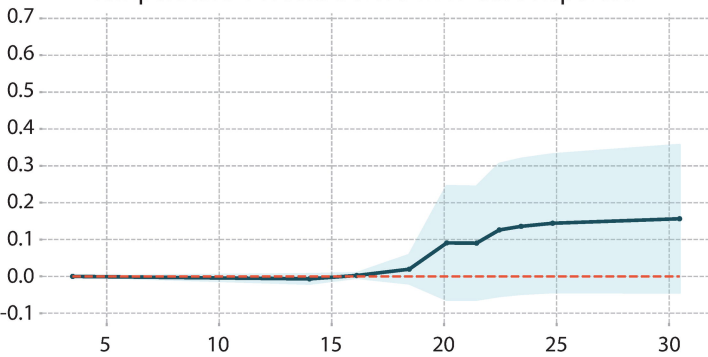
Total Population



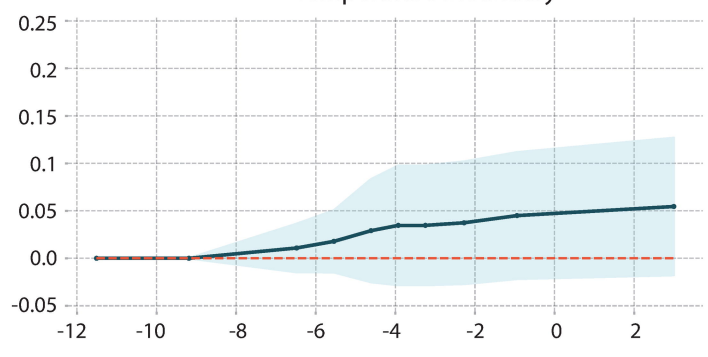
MIR mean



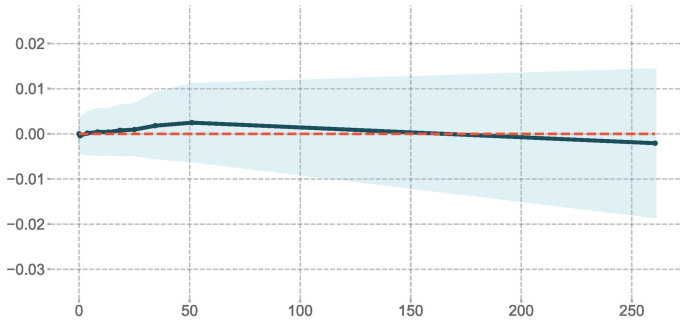
Temperature 4 Weeks before WNV Case Reported



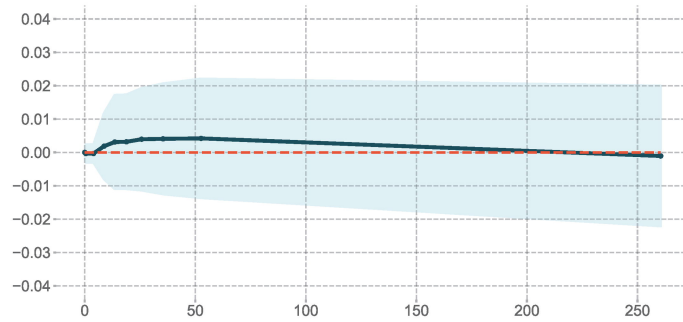
Temperature in January



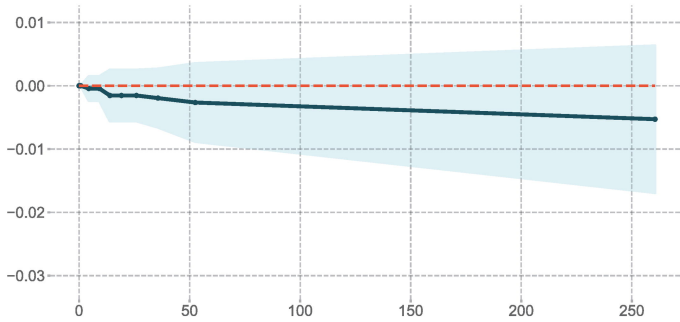
Precipitation of the Current Week



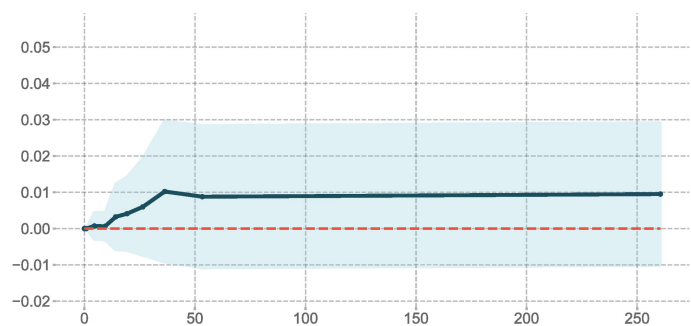
Precipitation 1 Week Before



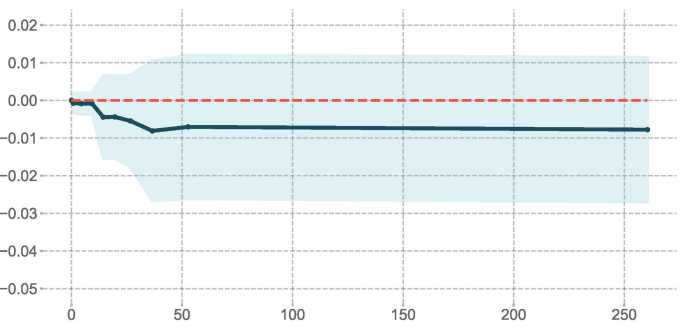
Precipitation 2 Weeks Before



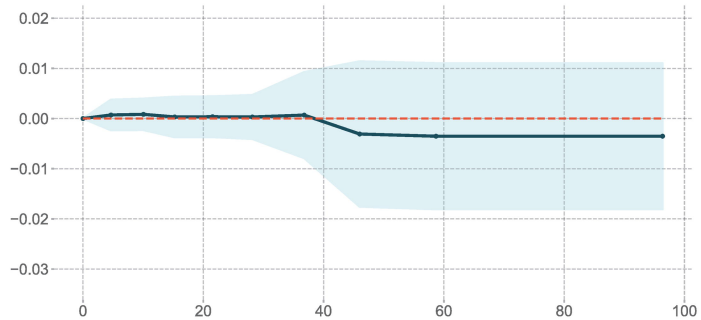
Precipitation 3 Weeks Before



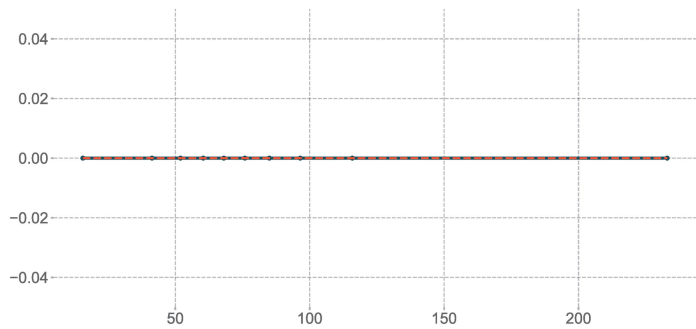
Precipitation 4 Weeks Before



Percentage of House after WWII



Income



Percentage of Land of Low Develop Intensity

