Identifying Predictors of Tree Mortality Post-Fire through Logistic Regression

Math 449: Categorical Data Analysis Project

Sequoia Andrade

May 2023

1 Introduction

Wildfires are becoming increasing frequent and severe, in part due to anthropogenic climate change. Each wildfire is unique, yet in some cases established trees survive post fire while other dies. To recognize which regions and forests are at higher risk of detrimental impacts due to wildfire, factors contributing to tree mortality after a fire should be identified. In this project, factors contributing to tree death post-fire are analyzed through a logistic regression with tree mortality as the target. Results indicate the tree species, fire intensity,

2 Data Set: Sequoia National Park

The data set used in this project is the "Sequoia and Yosemite National Parks Mortality and Fire Data (1990-2019) for Competition-Fire-Drought Interaction Analysis" data, abbreviated as SNFDP. The data set has a row for each individual tree and targets information relevant to tree mortality. The data set consists of 5674 rows of tree data spanning from 1990 to 2019. There are many variables, including fire year, forest plot, tree species, mortality date, tree diameter, crown scorch percentage (CVS), mortality after 3 years, mortality after 5 years, neighbors within 15m, neighbors within 5 meters, same species neighbors within 15 meters, same species neighbor within 5 meters, closed neighbor, the Hegyi competition index, and basal area of neighboring trees. Tree mortality after 3 years is considered the binary target, with 0 indicating no mortality and 1 indicating mortality. After an initial analysis, the fire year and plot are deemed multicollinear, with each plot only having one year of data, so the forest plot is removed from the model. The remaining predictor variables are tested for the full model.

3 Model Selection

3.1 Predictor Selection

The full model identifies fire year, species, diameter, CVS, neighbors within a 15 m distance, HEGYI, basal area of neighboring trees of the same species within a 15 m distance, and basal area of all neighboring trees within a 15 m distance as significant, as seen in Table 1. From here, predictor selection is performed in two ways. First, all non-significant predictors are removed and a new model is created. Next, a backwards step-wise selection is performed by starting at the full model.

From the manual removal of non-significant predictors, we get a model with all predictors except for the non-significant ones identified in the full model (i.e., fire year, species, diameter, CVS, neighbors within 15 m, hegyi competition, basal area of same species at 5 and 15 m, and basal area at 15m). The drop one likelihood ratio test is also used to confirm these variables are significant. From using the backwards stepwise selection function, we get a different model with the BA.8m predictor added in. This new model is significantly different from the manual model, with the manual model having a slightly larger deviance and

Table 1: Model coefficients from the full predictor set.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-97.1246573	32.9081144	-2.9513893	0.0031635
Fire_Year	0.0477459	0.0164195	2.9078714	0.0036390
SPECIESCADE	-0.1216292	0.2335422	-0.5208019	0.6025048
SPECIESPINUS	1.0724123	0.2357937	4.5480953	0.0000054
SPECIESQUKE	-0.3718897	0.6159572	-0.6037590	0.5460039
DBH	-0.0348047	0.0028969	-12.0143809	0.0000000
CVS	0.0560306	0.0019308	29.0193360	0.0000000
N.neighbors.15m	-0.0116724	0.0028092	-4.1551265	0.0000325
N.neighbors.5m	0.0149465	0.0169098	0.8838978	0.3767514
N.conspecific.neighbors.15m	0.0061092	0.0045479	1.3433104	0.1791715
N.conspecific.neighbors.5m	-0.0163792	0.0231974	-0.7060801	0.4801383
Nearest.neighbor	0.0148111	0.0618307	0.2395434	0.8106842
HEGYI	0.0099654	0.0023730	4.1995913	0.0000267
BA.conspecifics.5m	0.3790344	0.1918496	1.9756852	0.0481904
BA.conspecifics.15m	-0.1073306	0.0378634	-2.8346802	0.0045872
BA.3m	0.2420127	0.2001479	1.2091690	0.2265979
BA.5m	-0.1033754	0.1684204	-0.6137943	0.5393513
BA.8m	0.1636922	0.0813827	2.0113888	0.0442844
BA.11m	-0.0218021	0.0681153	-0.3200764	0.7489104
BA.15m	0.1122624	0.0516077	2.1753020	0.0296075
BA.20m	-0.0032577	0.0277613	-0.1173474	0.9065847

AIC. The final model chosen is the model from the step function, with the coefficient summary in Table 2 and the following variables: fire year, species, diameter, CVS, neighbors within 15 m, hegyi competition, basal area of same species at 5 and 15 m, and basal area at 8m and 15m.

Table 2: Final model coefficients.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-94.4780600	30.8408726	-3.063404	0.0021883
Fire_Year	0.0464870	0.0154101	3.016667	0.0025557
SPECIESCADE	-0.2426724	0.2117541	-1.146010	0.2517909
SPECIESPINUS	0.9022622	0.2007459	4.494549	0.0000070
SPECIESQUKE	-0.6186573	0.5941236	-1.041294	0.2977392
DBH	-0.0351114	0.0027965	-12.555669	0.0000000
CVS	0.0559842	0.0019040	29.402728	0.0000000
N.neighbors.15m	-0.0083344	0.0011646	-7.156139	0.0000000
HEGYI	0.0102746	0.0023364	4.397610	0.0000109
BA.conspecifics.5m	0.3440376	0.1345308	2.557314	0.0105484
BA.conspecifics.15m	-0.1027342	0.0359330	-2.859046	0.0042492
BA.8m	0.1477830	0.0639577	2.310638	0.0208529
BA.15m	0.0967892	0.0317793	3.045664	0.0023217

Model fit is evaluated by comparing the final model to the full model to test the null hypothesis that the models are significantly different. Results indicate the models are not significantly different, as shown in Table 3. Additionally, the model fit is evaluated by comparing the final model to the null model to test the hypothesis that the selected model is significantly different from the model. Results indicate the model is significantly different from the null model, as shown in Table 4.

The model fit can be examined using the residuals as seen in Figure 1, where the majority of the data is falling in a range with standard residuals less than 2. Overall the results indicate the model fits the data well with the selected predictors.

Table 3: Likelihood ratio test comparing final model to the full model

Resid. Df	Resid. Dev	Df	Deviance	Pr(¿Chi)
4527	1631.556	NA	NA	NA
4519	1627.751	8	3.804372	0.8743282

Table 4: Likelihood ratio test comparing final model to the null model

Resid. Df	Resid. Dev	Df	Deviance	Pr(¿Chi)
4527	1631.556	NA	NA	NA
4528	1636.861	-1	-5.30523	0.0212615

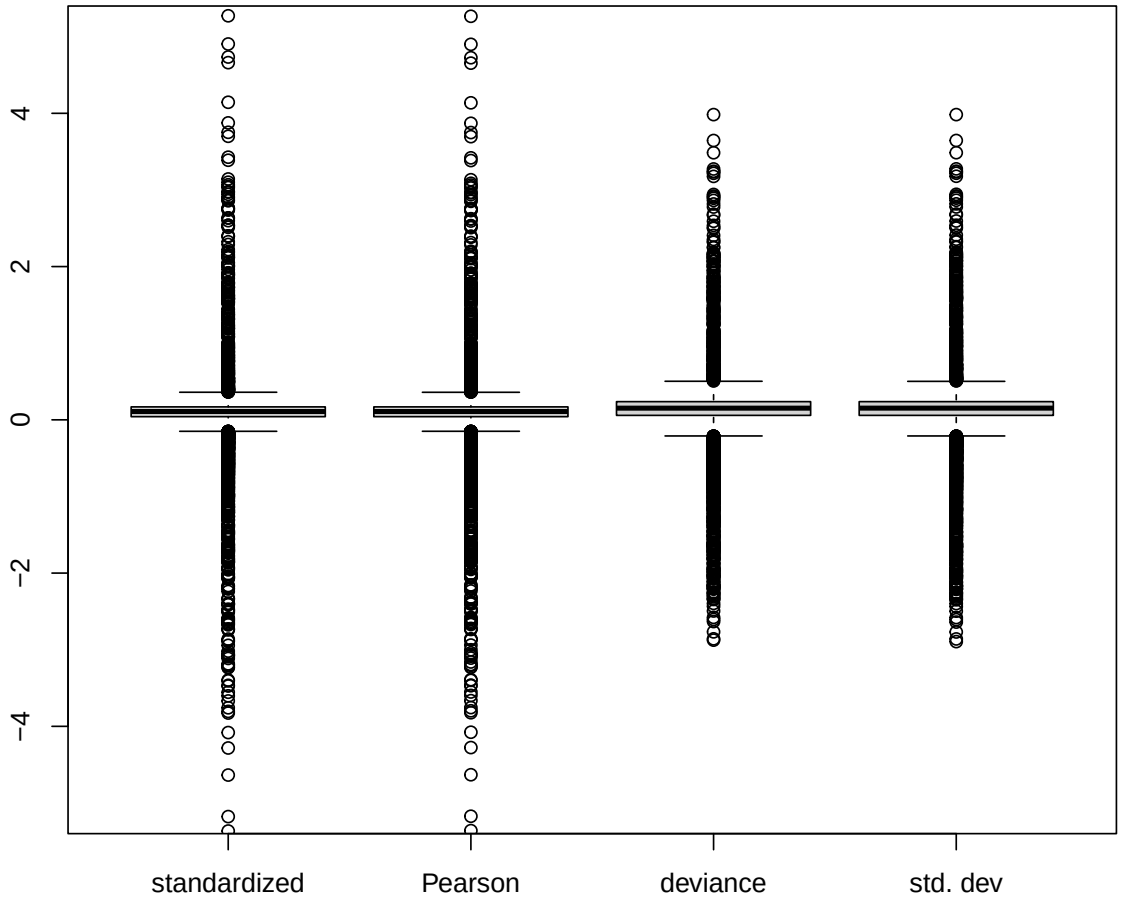

Figure 1: Residuals on the training set.

3.2 Inference

Inference for the multiplicative effect of the predictors is performed by finding the 95% confidence interval for each coefficient, as well as the confidence interval for the effect, shown in Table 5. From the inference

Table 5: Coefficient and multiplicative effective 95% confidence interval values.

	β	lower β	upper β	e^{β}	lower e^{β}	upper e^{β}
Fire_Year	0.0464870	0.0162839	0.0766901	1.0475845	1.0164172	1.0797075
SPECIESCADE	-0.2426724	-0.6577028	0.1723581	0.7845285	0.5180400	1.1881032
SPECIESPINUS	0.9022622	0.5088075	1.2957168	2.4651734	1.6633065	3.6536140
SPECIESQUKE	-0.6186573	-1.7831182	0.5458036	0.5386672	0.1681131	1.7259948
DBH	-0.0351114	-0.0405924	-0.0296304	0.9654979	0.9602205	0.9708042
CVS	0.0559842	0.0522523	0.0597160	1.0575809	1.0536415	1.0615351
N.neighbors.15m	-0.0083344	-0.0106171	-0.0060517	0.9917002	0.9894391	0.9939666
HEGYI	0.0102746	0.0056953	0.0148538	1.0103275	1.0057116	1.0149647
BA.conspecifics.5m	0.3440376	0.0803620	0.6077132	1.4106317	1.0836793	1.8362274
BA.conspecifics.15m	-0.1027342	-0.1731616	-0.0323067	0.9023668	0.8410017	0.9682096
BA.8m	0.1477830	0.0224283	0.2731377	1.1592613	1.0226817	1.3140812
BA.15m	0.0967892	0.0345028	0.1590756	1.1016281	1.0351050	1.1724266

table, we can see that the pinus species has the largest impact on mortality probability, with the odds of mortality for pinus trees 2.465 times the odds for non-pinus species. The basal area of the same species trees in a 15 m radius also has a large impact, with the odds of mortality increasing by 1.411 times for each one-unit increase in basal area. The CVS, a measure of fire intensity, has the next largest impact, the odds of mortality increasing by 1.058 times for each unit increase in crown scorch. Fire year also has a positive multiplicative effect, with the odds of mortality increasing by 1.048 for each year increase. The Hegyi competition metric has a small multiplicative effect, with each one-unit increase resulting in a 1.01 times increase in the odds of mortality. The remaining predictors have a negative coefficient, indicating increases in the predictor value result in decreases in the odds of mortality. Calocedrus decurrens (Cade) and Quercus kelloggii (QUKE) species have 0.785 and 0.546 odds of mortality respectively than trees that are neither. For each one centimeter increase in tree diameter, the odds of mortality decrease by 0.965 times. The number of neighbors in a 15m range results in a 0.992 times decrease in odds of mortality for each 1 neighbor increase. The basal area of same species trees with a 15 m radius decreases the odds of mortality by 0.902 times for each one-unit increase in basal area.

3.2.1 Model Visualization

Since the input space has multiple variables with a mix of categorical and continuous variables, the effect of the predictors is best visualized one plot at a time. Continuous predictor effects are shown in Figure 2, confirming the effects discussed in the inference section

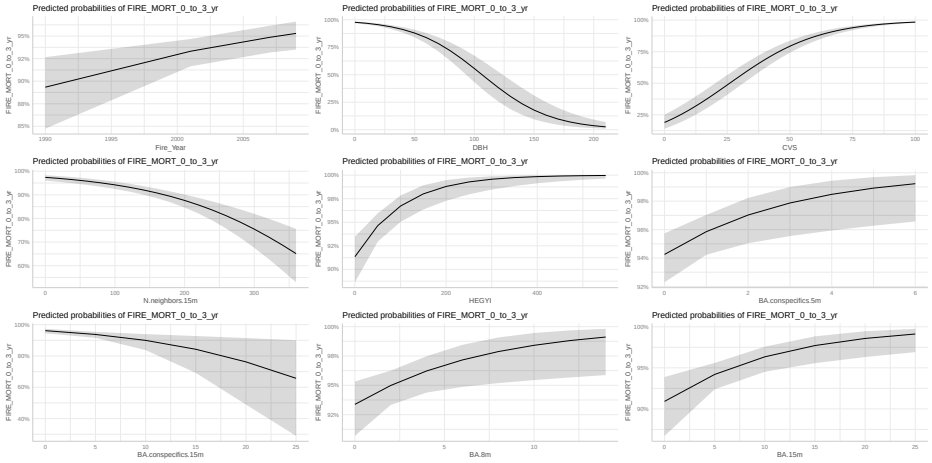

Figure 2: Visualizations of predictor effect for each continuous predictor.

3.3 Cutoff Value Selection

When $\pi_0 = 0.5$, the model has a specificity of 0.783, sensitivity of 0.969, and accuracy of 0.926. To identify if there is a better cutoff value, a range of values (0.25, 0.4, 0.5, 0.6, 0.75, 0.85, 0.9) are manually tested with metrics calculated. Results in Table 6 show a higher cutoff value results in better model performance. This is confirmed via the ROC curve in Figure 3, which identifies the best cutoff value of $\pi_0 = 0.912$. Additionally, the model fit is further validated with a high $AUC = 0.975$.

Table 6: Classification metrics on the validation set for different cutoff values.

p_0s	specificity	sensitivity	accuracy
0.25	0.6387833	0.9781860	0.8994709
0.40	0.7148289	0.9735936	0.9135802
0.50	0.7832700	0.9690011	0.9259259
0.60	0.8174905	0.9667049	0.9320988
0.75	0.8821293	0.9563720	0.9391534
0.85	0.9353612	0.9483352	0.9453263
0.90	0.9581749	0.9380023	0.9426808

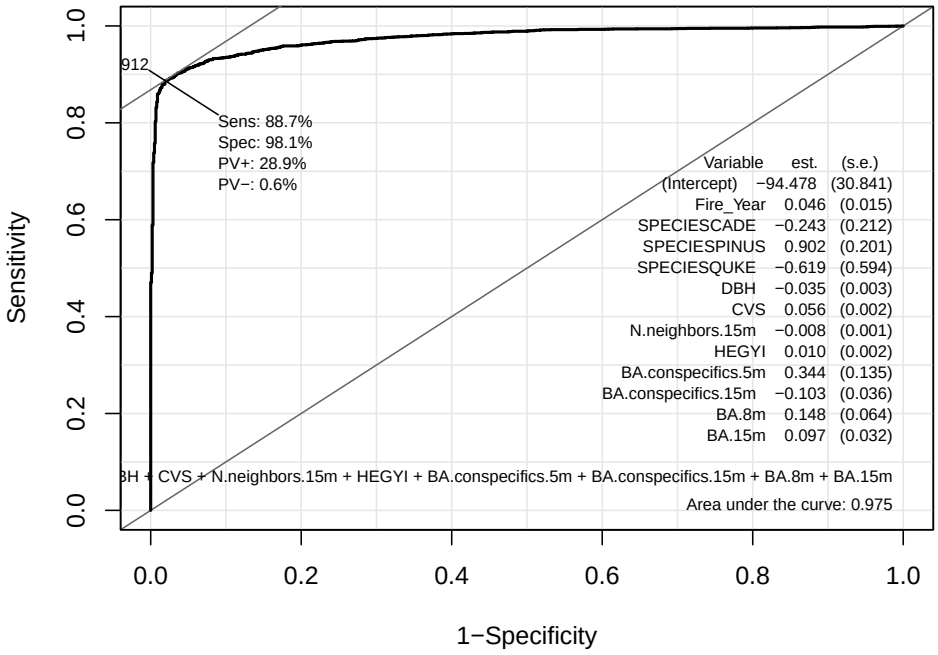

Figure 3: Receiver Operating Characteristics Curve for the final model.

3.4 Cross Validation

Leave-one-out and 10-fold cross-validation are both performed to estimate the model accuracy on the full dataset with a cutoff value of $\pi_0 = 0.5$. Both methods yeild an accuracy of 0.93.

3.5 Link Selection

Additional binomial models with identity and probit links are tested with the same input parameters and cutoff values. While the identity link model fails to converge, even with starting values given, the probit model outperforms the logit model with specificity, sensitivity, and accuracy slightly higher on the validation set, as seen in Table 7. The coefficients for the probit model are available in Table 8.

Table 7: Comparison of the logit and probit model on the validation set.

model	specificity	sensitivity	accuracy	AIC
logit	0.9581749	0.9322618	0.9382716	1657.6
probit	0.9619772	0.9311137	0.9382716	1647.7

Table 8: Coefficient values for the probit model

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-44.5760716	16.1439358	-2.761165	0.0057596
Fire_Year	0.0218487	0.0080668	2.708471	0.0067594
SPECIESCADE	-0.1348448	0.1054520	-1.278732	0.2009913
SPECIESPINUS	0.4191173	0.1061592	3.948006	0.0000788
SPECIESQUKE	-0.3749317	0.2966793	-1.263761	0.2063158
DBH	-0.0164245	0.0014432	-11.380460	0.0000000
CVS	0.0309714	0.0009362	33.081622	0.0000000
N.neighbors.15m	-0.0040402	0.0005936	-6.805811	0.0000000
HEGYI	0.0052671	0.0012174	4.326597	0.0000151
BA.conspecifics.5m	0.2002993	0.0716057	2.797255	0.0051539
BA.conspecifics.15m	-0.0618076	0.0192042	-3.218438	0.0012889
BA.8m	0.0671961	0.0341464	1.967880	0.0490819
BA.15m	0.0520497	0.0173299	3.003459	0.0026693

4 Conclusion

In this project, predictors of tree mortality after a fire in Sequoia National Park were analyzed via logistic regression. The best model included a combination of numeric and categorical predictors that consist of fire year, species, diameter, CVS, neighbors within 15 m, Hegyi competition, basal area of same species at 5 and 15 m, and basal area at 8m and 15m. The model classifies tree mortality very well using these predictors, with an accuracy of 0.938 on the validation set. Inference on the effect of the predictors identifies the tree species as the most impactful, followed by increased basal area, fire intensity, and fire year resulting in higher odds of mortality. In contrast, increases in tree diameter result in decreased odds of mortality. Hence, larger trees of certain species are more resilient to fires.