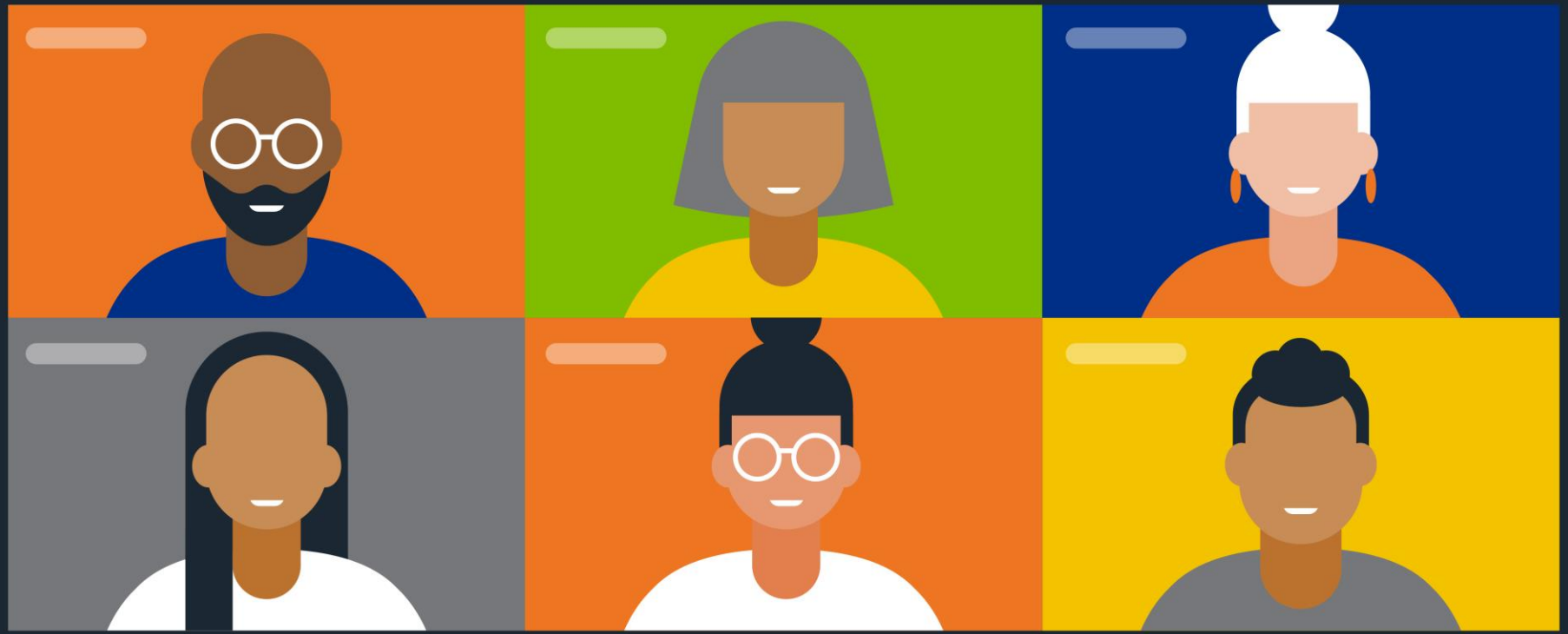


informs ANNUAL MEETING | 2020 VIRTUAL

Investigating Potential Bias And Discrimination In The Development of A Typical AI Platform For Heart Transplantation

Shuyu (Jade) Zhang, Zejian Wu, Chenhao You, Yan Shi, Hamid Ahady, Clark University

Introduction

- Heart failure is a global pandemic without cure
- Heart transplantation is the most effective treatment for patients with end-stage heart failure
- We want to help Decision makers have a predictive tool to facilitate their decision for organ matching
- We investigate the results of the latest researches in heart transplantation survival prediction for any evidences of bias in gender or region.

informs

Introduction

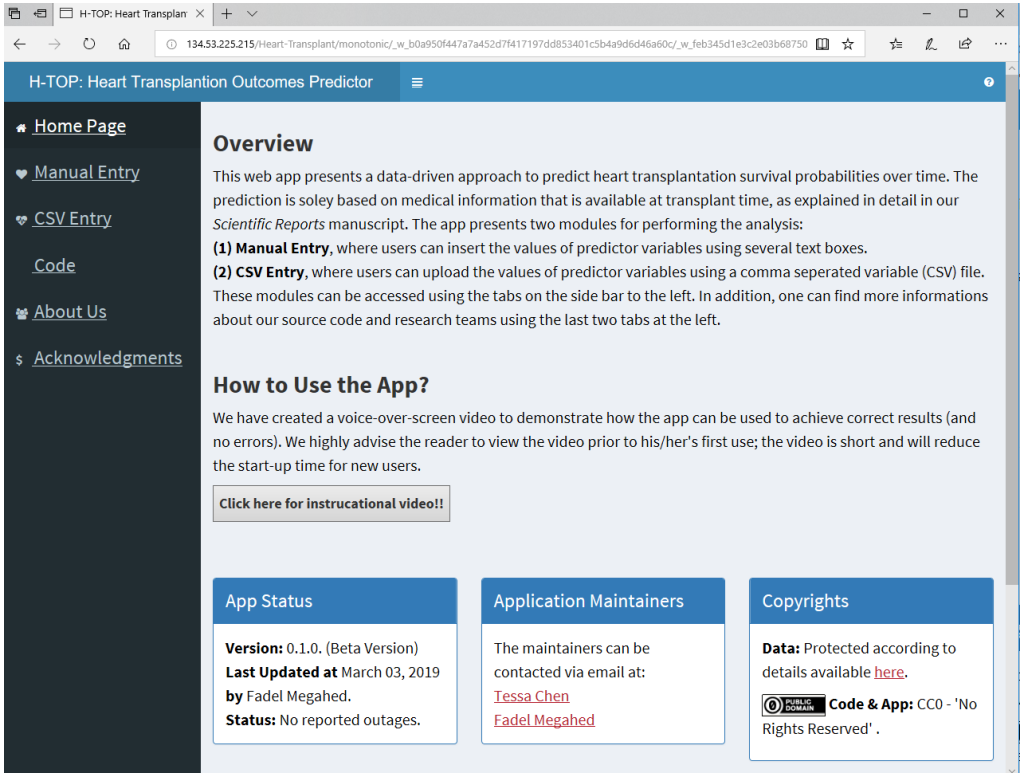
- Dataset: National registry of U.S. heart transplants from 1987-2016 (UNOS dataset)
- Survival of patients after one year from transplantation surgery is predicted.
- The best model (the highest AUC) is selected from all of the combinations:

Total No. of Data Mining Project Developed

	CRISP-DM Sections					
	Data Preparation			Training Model		
Factors	Categorical Imputation	Numerical Imputation	Encoding	Feature Selection	Resampling Method	Training Algorithm
No. of levels	4	2	2	3	5	9
Total Combinations	$4 \times 2 \times 2 \times 3 \times 5 \times 9 \times$ (5 fold cross validation) = 10,800					

- The training algorithm was Logistic Regression (the simplest one)

The survival tool



http://134.53.225.215/Heart-Transplant/monotonic/

Definition of Discrimination

- Initially originated from Latin for 'distinguishing’
- refers to an unjustified treatment of people based on belonging to some groups rather than their individual merits
- Human rights laws prohibit discrimination on the grounds of race, national or ethnic origin, color, religion, age, sex, sexual orientation, gender id entity or expression, marital status, family status, genetic characteristics, or disability.

Source: In 21st century Philippines, discrimination is still an inescapable way of life. GetRealPost

Discrimination in Machine Learning

- Discrimination due to algorithm is sometimes referred to as digital discrimination. In fact, digital discrimination could cause by biased dataset or the algorithm itself when sensitive attributes are included in the model

- Direct Discrimination
  - People that are similar in terms of non-protected characteristics should receive similar predictions
- Indirect Discrimination
  - Differences in predictions across groups of people can only be as large as justified by non-protected characteristics.

informs

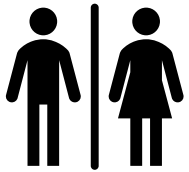
Protected Groups & Targets

- Gender of Patient
  - Male
  - Female
- Region
  - Southeast
  - Middle west
  - and Northeast

- Survival Status
  - 0

    The patient would not survive
  - 1

    The patient would survive
- Survival Possibility
  - From 0 to 1



We use statistical tests to investigate the existence of indirect discrimination in predicted survival status and survival possibility among gender and region

informs

Existence Tests

- **Regression Slope Test**
  - determine whether there is a significant linear relationship between an independent variable *X* and a dependent variable *Y*
  - $Y = B_0 + B_1X$
  - Hypothesis:

    $H_o: B_1 = 0$

    $H_a: B_1 \neq 0$

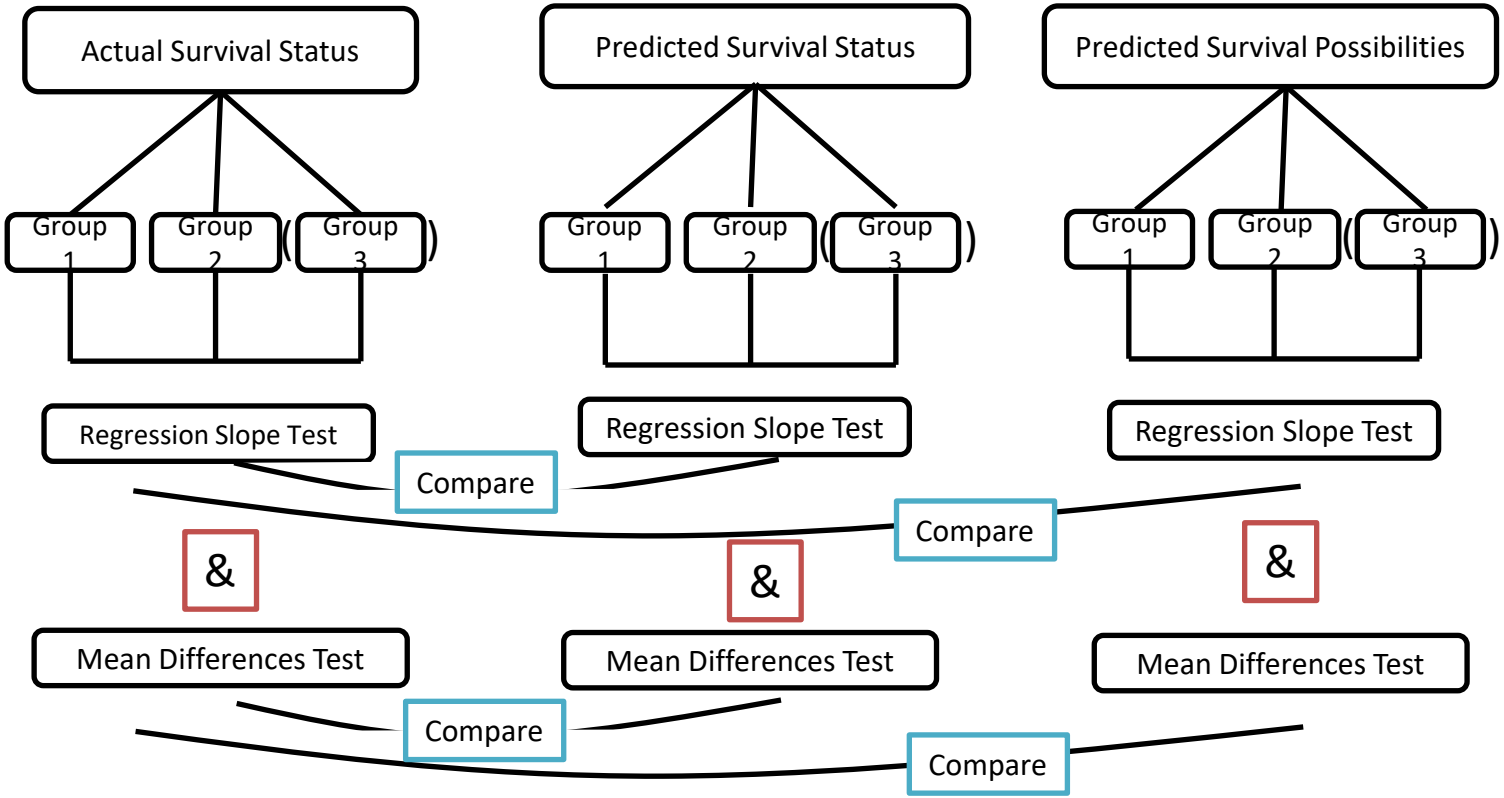
- **Mean Differences Test**
  - For two groups:
    - Ho : there is no difference between the two population means
    - Ha : there is difference between the two population means
  - For three groups:
    - ANOVA

informs

Methodology

- Gender
  1. Select the data generated by Logistic Regression
  2. Test if there is significant bias between male and female in the actual survival status (0 and 1)
  3. Test if there is significant bias between male and female in the predicted survival rate (0 and 1)
  4. Test if there is significant bias between male and female in the predicted survival possibility (0 to 1)
  5. Compare the test results
- Region
  - Similar process
  - Detect the bias between Southeast, Middle west, and Northeast

Methodology



informs

Result: Gender – Regression Test

$Y = B_0 + B_1X$

$H_o: B_1 = 0$

Actual Survival Status				
Year	slope	std_error	p-value	conclusion
0	0.001511	0.006605	0.819071	Accept
1	0.012958	0.009701	0.181679	Accept
2	0.002992	0.011007	0.785788	Accept
3	0.013493	0.011838	0.254377	Accept
4	0.013493	0.011838	0.254377	Accept
5	0.012717	0.013391	0.342328	Accept
6	0.022473	0.014057	0.109933	Accept
7	0.00921	0.014482	0.524829	Accept
8	0.001508	0.014928	0.919531	Accept
9	-0.00653	0.015153	0.666347	Accept
10	-0.01831	0.015338	0.232594	Accept

Predicted Survival Status				
Year	slope	std_error	p-value	conclusion
0	0.050847	0.011930913	2.05E-05	Reject
1	0.035387	0.012385337	0.004284322	Reject
2	0.038522	0.01273741	0.002499292	Reject
3	0.045311	0.013024613	0.00050617	Reject
4	0.045311	0.013024613	0.00050617	Reject
5	0.089718	0.013694132	6.08E-11	Reject
6	0.077486	0.014078576	3.85E-08	Reject
7	0.052496	0.014398259	0.000268411	Reject
8	0.047808	0.014736314	0.001183829	Reject
9	0.034139	0.014957594	0.022501932	Reject
10	0.042433	0.015247362	0.005404143	Reject

Predicted Survival Possibilities				
Year	slope	std_error	p-value	conclusion
0	0.019286	0.004130946	3.07E-06	Reject
1	0.01465	0.003554636	3.80E-05	Reject
2	0.011839	0.003172059	0.000190972	Reject
3	0.014741	0.003082325	1.76E-06	Reject
4	0.020767	0.003106283	2.46E-11	Reject
5	0.028701	0.003402421	3.94E-17	Reject
6	0.021556	3.41E-10	0.003428245	Reject
7	0.018114	0.003742111	1.32E-06	Reject
8	0.017326	0.0045023	0.000120078	Reject
9	0.013855	0.005581601	0.013079482	Reject
10	0.01545	0.00665848	0.020357182	Reject

Result: Gender – Mean Differences Test

Ho: there is no difference between the two population means

Actual Survival Status			
Year	statistic	p-value	conclusion
0	0.228746	0.819071	Accept
1	1.119031	0.263173	Accept
2	0.271792	0.785788	Accept
3	1.139863	0.254377	Accept
4	1.575005	0.115296	Accept
5	0.949638	0.342328	Accept
6	1.598701	0.109933	Accept
7	0.635953	0.524829	Accept
8	0.101028	0.919531	Accept
9	-0.43119	0.666347	Accept
10	-1.19383	0.232594	Accept

Predicted Survival Status			
Year	statistic	p-value	conclusion
0	4.261827	2.05E-05	Reject
1	1.760658	0.078348741	Accept
2	3.024326	0.002499292	Reject
3	3.478876	0.00050617	Reject
4	5.693438	1.29E-08	Reject
5	6.551596	6.08E-11	Reject
6	5.503812	3.85E-08	Reject
7	3.645984	0.000268411	Reject
8	3.244222	0.001183829	Reject
9	2.282371	0.022501932	Reject
10	2.782945	0.005404143	Reject

Predicted Survival Possibilities			
Year	statistic	p-value	conclusion
0	4.668684	3.07E-06	Reject
1	2.729504	0.006362	Reject
2	3.732301	0.000191	Reject
3	4.782497	1.76E-06	Reject
4	6.68547	2.46E-11	Reject
5	8.435379	3.94E-17	Reject
6	6.287807	3.41E-10	Reject
7	4.840622	1.32E-06	Reject
8	3.848365	0.00012	Reject
9	2.482331	0.013079	Reject
10	2.320334	0.020357	Reject

Result: Region – Regression Test

REGION_MIDWEST

Actual Survival Status

Year	slope	p-value	std_error	conclusion
0	-0.00575	0.376377	0.006504	Accept
1	0.011935	0.208335	0.009486	Accept
2	0.016822	0.117115	0.010734	Accept
3	0.015116	0.19611	0.011692	Accept
4	0.023837	0.055766	0.01246	Accept
5	0.021162	0.105789	0.013082	Accept
6	0.021986	0.108197	0.013685	Accept
7	0.023312	0.100971	0.014211	Accept
8	0.032278	0.026408	0.014535	Reject
9	0.021005	0.156257	0.014814	Accept
10	0.019313	0.198195	0.015008	Accept

Predicted Survival Status

Year	slope	p-value	std_error	conclusion
0	0.058056	7.84E-07	0.011746	Reject
1	0.054162	7.71E-06	0.012102	Reject
2	0.072223	6.03E-09	0.012406	Reject
3	0.088542	5.70E-12	0.012837	Reject
4	0.122215	1.09E-20	0.013067	Reject
5	0.14526	1.63E-27	0.013311	Reject
6	0.101202	1.56E-13	0.013682	Reject
7	0.119321	2.73E-17	0.01407	Reject
8	0.100935	1.93E-12	0.01431	Reject
9	0.122428	4.80E-17	0.014547	Reject
10	0.120963	4.48E-16	0.014845	Reject

Predicted Survival Probabilities

Year	slope	p-value	std_error	conclusion
0	0.021012	2.44E-07	0.004067	Reject
1	0.019958	9.35E-09	0.003472	Reject
2	0.025133	4.22E-16	0.003084	Reject
3	0.024547	7.21E-16	0.003037	Reject
4	0.031309	3.71E-25	0.003011	Reject
5	0.041423	1.14E-35	0.003305	Reject
6	0.030581	5.01E-20	0.003326	Reject
7	0.035749	1.83E-22	0.003653	Reject
8	0.036869	3.75E-17	0.004366	Reject
9	0.047995	1.17E-18	0.005425	Reject
10	0.051314	2.95E-15	0.006484	Reject

REGION_NOTH_EAST

Actual Survival Status

Year	slope	p-value	std_error	conclusion
0	-1.31E-02	5.07E-02	6.70E-03	Accept
1	-3.85E-02	8.69E-05	9.81E-03	Accept
2	-3.27E-02	3.35E-03	1.12E-02	Reject
3	-3.23E-02	7.27E-03	1.20E-02	Reject
4	-1.94E-02	1.32E-01	1.29E-02	Accept
5	-2.76E-02	4.14E-02	1.36E-02	Reject
6	-2.00E-02	1.58E-01	1.42E-02	Accept
7	-2.26E-02	1.20E-01	1.46E-02	Accept
8	-2.25E-02	1.32E-01	1.49E-02	Accept
9	-2.03E-02	1.84E-01	1.53E-02	Accept
10	-1.96E-02	2.05E-01	1.54E-02	Accept

Predicted Survival Status

Year	slope	p-value	std_error	conclusion
0	-1.19E-01	7.61E-23	1.21E-02	Reject
1	-9.37E-02	7.21E-14	1.25E-02	Reject
2	-4.53E-02	4.61E-04	1.29E-02	Reject
3	-5.03E-02	1.48E-04	1.32E-02	Reject
4	-5.14E-02	1.47E-04	1.35E-02	Reject
5	-9.82E-02	1.55E-12	1.39E-02	Reject
6	-4.76E-02	8.16E-04	1.42E-02	Reject
7	-3.09E-02	3.31E-02	1.45E-02	Reject
8	-3.04E-02	3.95E-02	1.47E-02	Reject
9	-8.42E-02	2.27E-08	1.50E-02	Reject
10	-6.97E-02	5.58E-06	1.53E-02	Reject

Predicted Survival Probabilities

Year	slope	p-value	std_error	conclusion
0	-4.96E-02	1.64E-32	4.16E-03	Reject
1	-3.43E-02	1.11E-21	0.003581	Reject
2	-1.45E-02	6.62E-06	3.22E-03	Reject
3	-1.69E-02	7.23E-08	3.13E-03	Reject
4	-1.64E-02	1.55E-07	3.12E-03	Reject
5	-3.08E-02	4.65E-19	3.44E-03	Reject
6	-1.84E-02	1.13E-07	3.46E-03	Reject
7	-1.20E-02	1.49E-03	3.77E-03	Reject
8	-1.71E-02	1.43E-04	4.50E-03	Reject
9	-3.63E-02	1.06E-10	5.61E-03	Reject
10	-3.66E-02	4.46E-08	6.68E-03	Reject

REGION_SOUTH_EAST

Actual Survival Status

Year	slope	p-value	std_error	conclusion
0	-2.74E-03	6.43E-01	5.90E-03	Accept
1	-0.00921	0.288	0.008672	Accept
2	-0.00779	0.427971	0.009823	Accept
3	-0.00766	0.470637	0.010611	Accept
4	-0.02094	0.065238	0.011356	Accept
5	-0.01105	0.354125	0.011925	Accept
6	-0.0241	0.054279	0.012519	Accept
7	-0.00837	0.516715	0.012909	Accept
8	-0.00628	0.635476	0.013246	Accept
9	-0.01318	0.328455	0.013485	Accept
10	-0.01504	0.27007	0.013638	Accept

Predicted Survival Status

Year	slope	p-value	std_error	conclusion
0	-1.45E-02	1.76E-01	1.07E-02	Accept
1	-0.02803	0.011354747	0.011071	Reject
2	-0.08361	1.81E-13	0.012914	Reject
3	-0.09059	8.00E-15	0.01164	Reject
4	-0.12751	1.19E-26	0.011888	Reject
5	-0.12473	1.40E-24	0.012143	Reject
6	-0.10408	1.02E-16	0.012504	Reject
7	-0.12768	1.96E-23	0.012752	Reject
8	-0.106	4.65E-16	0.013019	Reject
9	-0.08311	4.11E-10	0.013275	Reject
10	-0.08427	4.90E-10	0.013521	Reject

Predicted Survival Probabilities

Year	slope	p-value	std_error	conclusion
0	-6.04E-03	1.02E-01	0.003696	Accept
1	-0.01148	0.000306026	0.003178	Reject
2	-0.0256	1.35E-19	0.00282	Reject
3	-0.02509	9.74E-20	0.002753	Reject
4	-0.0339	6.44E-35	0.002736	Reject
5	-0.03618	7.28E-33	0.003015	Reject
6	-0.02974	1.92E-22	0.003041	Reject
7	-0.03728	3.73E-29	0.00331	Reject
8	-0.03405	1.39E-17	0.003977	Reject
9	-0.02777	2.21E-08	0.004957	Reject
10	-0.02714	4.51E-06	0.005912	Reject

informs

Result: Region – ANOVA

Ho: there is no difference in means

Actual Survival Status			
Year	statistic	p-value	conclusion
0	0.576839036	0.561695	Accept
1	4.252878352	0.014277	Reject
2	3.886647697	0.020558	Reject
3	3.046513543	0.047591	Reject
4	3.071013217	0.046444	Reject
5	2.648017499	0.070874	Accept
6	2.553425685	0.077904	Accept
7	2.042358654	0.129822	Accept
8	2.794213227	0.061255	Accept
9	1.651012699	0.191961	Accept
10	1.492085417	0.225008	Accept

Predicted Survival Status			
Year	statistic	p-value	conclusion
0	42.93389	2.859E-19	Reject
1	23.49655	7.005E-11	Reject
2	31.68267	2.005E-14	Reject
3	40.23185	4.294E-18	Reject
4	72.41616	8.078E-32	Reject
5	93.52239	1.008E-40	Reject
6	44.89891	4.496E-20	Reject
7	59.53555	2.652E-26	Reject
8	40.39344	3.913E-18	Reject
9	49.93881	3.366E-22	Reject
10	44.6354	6.22E-20	Reject

Predicted Survival Possibilities			
Year	statistic	p-value	conclusion
0	58.41876	6.57133E-26	Reject
1	36.53723	1.78681E-16	Reject
2	55.54145	1.16897E-24	Reject
3	57.39094	1.94586E-25	Reject
4	95.55743	1.31515E-41	Reject
5	129.3908	9.61627E-56	Reject
6	69.33773	1.76977E-30	Reject
7	78.31291	2.95523E-34	Reject
8	54.35979	4.3223E-24	Reject
9	55.54211	1.39396E-24	Reject
10	42.37661	5.71776E-19	Reject

informs

Conclusion

- Existence of Gender Bias
- Existence of Region Bias
- The highest performing Model was in favor of female
- The highest performing Model was in favor of region northeast
- Any scoring algorithm based on the produced tool is prone to bias and needed to be optimized to reduce bias

Future Studies

- Measurement of bias
- Mathematical improvements in algorithms: developing cost function for bias
- Run algorithm separately for different groups to investigate if there exists overfitting/underfitting problem for some specific groups
- Sensitivity analysis of accuracy vs. bias

Thank you!