Developing a Titanic survival scorecard: Risk analysis of populations through statistical scoring methods

Dominic Vincent Ligot, CirroLytix Research Services

23 January 2022

Abstract

Objectives. We developed a survival scorecard from the Titanic passenger dataset using statistical scoring methods. We discuss the scorecard development process, assess the effectiveness of statistical scorecards, and analyze the characteristics of Titanic passengers that led to survival.

Methods. From the Titanic dataset of 1,309 passengers and a binary dependent variable representing survival, we assessed nine (9) features using chi-square, Weight of Evidence (WoE), and Information Value. A logistic regression was fitted on the feature WoEs to predict survival, feature coefficients were used to determine score weights for each attribute, and an additive model on attribute scores per passenger determined survival scores. The resulting scorecards were assessed for risk ranking, and the characteristics of the passenger population were assessed for survivability and population shifts.

Results. The resulting survival scorecard was able to rank survivability amongst Titanic passengers (K-S 0.558, ROC 0.892, CAP 0.765) with survival classification accuracy varying by score cutoff (AR 0.63 – 0.84). Sex was the strongest predictor of survivability (IV 145), followed by fare amount (IV 53), cabin class (IV 52), and passenger class (IV 51). Women passengers had four times higher survivability compared to men (72% vs. 16%). Passengers who paid $100 or more for their trip had nearly ten times higher survivability compared to free passengers (75% vs. 7.6%). Cabin passengers had higher survivability compared to non-cabin passengers with cabin B having nearly three times higher survivability compared to non-cabin passengers (75% vs. 27%). Class 1 passengers had nearly three times higher survivability compared to Class 3 passengers (62% vs. 22%).

Conclusion. This paper illustrates the benefits of statistical scoring methods compared to other machine learning approaches in the analysis of event likelihood risks and performing population risk segmentation. Machine learning approaches usually focus on prediction accuracy while scorecards allow for cross-sectional analysis of population risks. Scorecard cut-offs provide avenues for decision-making on populations accounting for tradeoffs in accuracy, recall, precision, and specificity.

Keywords: credit scoring, logistic regression, weight of evidence, information value, kolmogorov-smirnov, receiver operating characteristic, cumulative accuracy profile, lorenz curve, population stability index, titanic

JEL Classification: C44, D81, R23

1. Background

The Titanic passenger dataset has been analyzed and used as a teaching tool for statistics and mixed methods research (Lindemann & Stolz, 2021). Past analysis of the Titanic dataset has established the impact of sex and social class on survival, survival declined with social class and women had a higher rate of survival to men (Hall, 1986).

Machine learning methods have been used to predict survivability on the Titanic with a typical focus on classification accuracy (Farag & Hassan, 2018). Statistical scoring is a predictive modeling technique used widely in the financial industry for the assessment of credit risk of borrowers. While machine learning for classification outputs a class prediction or actual class probabilities for a given member of a population, in contrast, scoring aims to generate an abstracted value or score that represents statistical odds or probabilities for a given outcome – say loan or credit card default. Scoring is a form of interpretable machine learning, similar to linear and logistic regression, but with model coefficients scaled and adjusted into more intuitive score values which can be easier to interpret and explain to non-technical audiences. If accurate, a scorecard will be able to rank the risk of an outcome using score values, and users can set a defined threshold or "score cut-off" at which to make a decision – say approve or decline a loan application (Siddiqi, 2017).

Scoring's primary use-case is credit risk (Skantzos & Castelein, 2016), and while machine learning methodologies have been used to analyze the Titanic dataset, the application of scoring on non-financial datasets is rare (Chaiyadecha, 2020).

In this paper, we developed a survival scorecard from the Titanic passenger dataset using statistical scoring methods. We will discuss the scorecard development process applied to a non-financial use-case, and demonstrate the applicability of scoring methods as a tool for population risk analysis.

2. Related Work

The Titanic dataset has been a popular target to illustrate statistical tests and machine learning models. Given a binary outcome for survivability and various explanatory variables, the dataset can be used to demonstrate hypothesis testing (Dixon & Griffiths, 2004) and is a case study for logistic regression (Harrell, 2001).

Apart from logistic regression, machine learning classification algorithms have been used to predict survival on the titanic such as Decision Tree, Random Forest, Naïve Bayes (Nair, 2017), as well as gradient boosting methods, XGBoost and Catboost. Methods are typically benchmarked using classification accuracy rates via confusion matrix and area under the curve (AUC) calculations using the Receiver Operating Characteristic (ROC) calculation (Ibrahim & Abdulaziz, 2020). Feature generation through hierarchical and K-Means clustering has also been used to segment the population of passengers into risk segments (Nikitina & Zamnius, 2019).

3. Scorecard Development Method

3.1 Titanic Dataset

There are several openly published sources of Titanic passenger data. The most popular is the Kaggle dataset (Kaggle, 2020) used in predictive model development competitions. This dataset is similar and is likely copied from an open dataset made available by Thomas Cason (Cason &

Harrell, 2017) and accessible through the R library. Another well-maintained open dataset is from a mixed-methods demonstration paper (Lindemann & Stolz, 2021). Both datasets trace their source to Encyclopedia Titanica, a website dedicated to facts and history of the RMS Titanic (Encylopedia Titanica, 1996-2021) and also offers a comprehensive dataset to its premium members.

The Kaggle/Cason dataset has 1,309 records, while the Lindemann/Stolz dataset has 2,207 records for quantitative analysis as well as additional datasets for qualitative data. We have chosen to use the Cason dataset for this paper due to the presence of additional features of interest (e.g. fare, cabin, embarkation, and destination) that can be used for scoring.

3.1.1 Train-Test-Validate Sampling

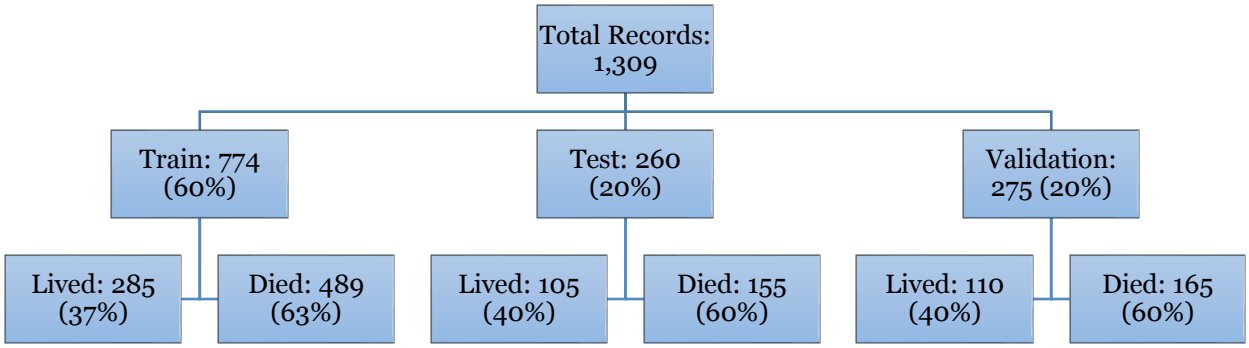

Figure 1: Population Sampling

From the initial population of 1,309 records, the dataset was randomly split into Train, Test, and Validation sets. Each set had roughly same proportions of surviving and perished passengers. The train dataset was used to develop the scoring models but the test and validate data were used to confirm the scorecard's accuracy and perform the population risk analysis.

3.2 Calculating Scores

3.2.1 Feature Generation

Continuous variables (e.g. Age, Fare, Siblings, Parents) were binned into categories. Categories that were too sparse (e.g. <1% of total) were grouped with other sparse categories. Blanks were retained as a distinct category.

3.2.2 Survival Rate, Weight of Evidence

Survival rates were calculated for each categorical attribute (Survival = Lived / Total). Weight of Evidence was also calculated.

Equation 1: Weight of Evidence Calculation

$$WOE = \ln\left(\frac{distribution\ of\ Lived}{distribution\ of\ Died}\right) \times 100$$

3.2.3 Feature Importance, Chi-square, Information Value

Features were ranked using the chi-square test of independence and information value. Features with chi-square p-values greater than 0.5 or information values less than 2.0 would be candidates for removal (Siddiqi, 2017).

Equation 2: Information Value Calculation

$$Information\ Value = (distribution\ of\ Lived - distribution\ of\ Died) \times WOE$$

Table 1: Information Value Strength

Information Value Range	Indication
Less than 2.0	Not useful
2.0 to 10.0	Weak power
10.0 to 30.0	Medium power
30.0 to 50.0	Strong power
Greater than 50.0	Very strong (possible over fit)

3.2.4 Logistic Regression and Score Weights

Once the final feature list was obtained, the dataset would be recoded to reflect the WOE values in place of the original categories. A logistic regression was fit on the WOE values as independent variables with survival as the target dependent variable to generate the regression coefficients and intercept (Skantzos & Castelein, 2016).

We assigned a value of 20 points to double the odds of survival and a target score of 600 corresponding to a survival odds of 30:1 (Chaiyadecha, 2020). Scores were then scaled using a combination of the Factor, Offset, logistic regression coefficients and intercept.

The resulting scores were rounded off for ease of use. Once score weights were determined for each attribute, an additive model combines this to generate a score per passenger.

Equation 3: Calculating Score Weights and Scaling (i = features, j = attributes)

$$Points\ to\ double\ odds\ (pdo) = 20$$

$$factor = \frac{pdo}{\ln(2)}$$

$$offset = 600 - (factor \times \ln(30))$$

$$Score_i = \sum_{i=1}^{n} \left(\frac{offset}{n} + factor \times (WOE_{ij} \times \beta_i + \frac{\beta_0}{n}) \right)$$

4. Results

4.1 Feature Importance

Based on Chi-square and IV, all nine features were retained for modeling. Sex was the strongest predictor of survivability (IV 145), followed by fare amount (IV 53), cabin class (IV 52), and passenger class (IV 51).

Table 2: Feature Importance and Regression Results

Feature	X2	d.f.	p	IV	b	b*
Age	10.61	5	0.0596	5.85	0.0073	0.0075
Cabin Class	94.67	6	< 0.0001	52.50	0.0052	0.0050
Fare	89.50	6	< 0.0001	53.60	-0.0033	
Destination	51.83	11	< 0.0001	30.34	0.0073	0.0071
Sex	242.73	1	< 0.0001	145.21	0.0110	0.0106
Siblings Spouses	26.95	4	< 0.0001	19.86	0.0046	0.0043
Parents Children	34.92	3	< 0.0001	18.97	0.0069	0.0055
Embarkation	28.23	2	< 0.0001	15.14	0.0065	0.0058
Passenger Class	91.55	2	< 0.0001	51.87	0.0071	0.0049
Intercept					-0.5240	-0.5206

b* - alternate model omitting Fare (see 4.3)

4.2 Survival Rate, Weight of Evidence, Score Weights

Women passengers had four times higher survivability compared to men (72% vs. 16%). Passengers who paid $100 or more for their trip had nearly ten times higher survivability compared to free passengers (75% vs. 7.6%). Cabin passengers had higher survivability compared to non-cabin passengers with cabin B having nearly three times higher survivability compared to non-cabin passengers (75% vs. 27%). Class 1 passengers had nearly three times higher survivability compared to Class 3 passengers (62% vs. 22%).

Survivability generally increased with age but was not consistent; 51-60 year olds had the highest survivability however, 61+ had the lowest. Survivability by destination was mixed, PQ, MI had the highest survivability. For Siblings and Spouses, survivability increased until 2 and declined at 3 or more. Passengers with 1-2 Parents and Children had the best survivability while 0 and 3 or more had lower survival. Passengers who embarked from Cherbourg had the best survivability. The resulting score weights assigned higher values for higher survivability with the exception of fare which reversed the weight order (i.e. higher values for lower survivability) due to the negative logistic regression coefficient.

4.3 Feature Correlation

Using Weight of Evidence values, Cabin Class, Fare, and Passenger Class were highly correlated (R 0.53 – 0.69) which suggests one of these features could be dropped without diminishing the power of the scorecard. This explains Fare having a negative coefficient in the model resulting in a reverse ordered score weight for its attributes (Ranganathan, Pramesh, & Aggarwal, 2017). We ran an alternate regression omitting Fare from the model and resulted with all positive coefficients (b*). For the remainder of the paper we retained the original regression result to retain as much information within the model, and the resulting scorecard performed well nonetheless.

Table 3: Feature Attribute Summary Counts (Train)

Feature	Attribute	Died	Lived	Total	Survival	WOE	Score
Age	0-17	155	87	242	0.36	-3.8	53
	18-30	185	96	281	0.34	-11.6	52
	31-40	68	55	123	0.45	32.8	61
	41-50	51	26	77	0.34	-13.4	51
	51-60	13	16	29	0.55	74.8	70
	61+	17	5	22	0.23	-68.4	40
Cabin	Blank	429	164	593	0.28	-42.2	48
	A	8	11	19	0.58	85.8	67
	B	10	31	41	0.76	167.1	79
	C	20	34	54	0.63	107.1	70
	D	9	14	23	0.61	98.2	69
	E	8	21	29	0.72	150.5	77
	F, G, T	5	10	15	0.67	123.3	73
Fare*	0	16	2	18	0.11	-194.5	73
	0-10	366	108	474	0.23	-82.8	62
	11-20	162	99	261	0.38	-9.6	55
	21-30	118	95	213	0.45	40.2	50
	31-40	46	33	79	0.42	23.0	52
	41-100	77	103	180	0.57	70.5	47
	100+	24	60	84	0.71	163.8	38
Destination	Blank	303	126	429	0.29	-33.8	47
	Others	48	42	90	0.47	40.6	63
	IL	17	5	22	0.23	-68.4	40
	MA	13	6	19	0.32	-23.3	49
	MI	7	14	21	0.67	123.3	80
	MN	9	2	11	0.18	-96.4	34
	NJ	12	11	23	0.48	45.3	64
	NY	52	54	106	0.51	57.8	66
	OH	6	8	14	0.57	82.8	71
	ON	13	1	14	0.07	-202.5	12
	PA	6	10	16	0.63	105.1	76
	PQ	3	6	9	0.67	123.3	80
Sex	female	77	204	281	0.73	151.4	102
	male	412	81	493	0.16	-108.7	20
Siblings Spouses	0	348	175	523	0.33	-14.8	52
	1	97	92	189	0.49	48.7	61
	2	16	15	31	0.48	47.5	60
	3	5	2	7	0.29	-37.6	49
	4, 5, 8	23	1	24	0.04	-259.6	20
Parents Children	0	402	188	590	0.32	-22.0	50
	1	45	58	103	0.56	79.4	70
	2	29	36	65	0.55	75.6	69
	3, 4, 5, 6, 9	13	3	16	0.19	-92.6	36
Embarkation	C	74	89	163	0.55	72.4	68
	Q	50	26	76	0.34	-11.4	52
	S, blank	365	170	535	0.32	-22.4	50
Passenger Class	1	74	121	195	0.62	103.2	75
	2	90	68	158	0.43	26.0	59
	3	325	96	421	0.23	-68.0	40

*Fare has reverse ordered score weights (see 4.3)

Table 4: WoE Feature Correlation (Train)

	Age	Cabin	Fare	Destination	Sex	Siblings	Parents	Embarkation	Passenger
Age		0.126	0.106	0.064	0.031	0.036	0.036	0.030	0.138
Cabin Class*	0.126		0.534	0.232	0.154	0.159	0.109	0.252	0.699
Fare*	0.106	0.534		0.310	0.274	0.093	0.322	0.260	0.695
Destination	0.064	0.232	0.310		0.087	0.137	0.160	0.043	0.373
Sex	0.031	0.154	0.274	0.087		0.109	0.194	0.086	0.121
Siblings	0.036	0.159	0.093	0.137	0.109		-0.164	0.125	0.200
Parents	0.036	0.109	0.322	0.160	0.194	-0.164		0.042	0.077
Embarkation	0.030	0.252	0.260	0.043	0.086	0.125	0.042		0.260
Pass. Class*	0.138	0.699	0.695	0.373	0.121	0.200	0.077	0.260	

*Cabin Class, Fare, and Passenger Class have high correlation amongst the features.

4.4 Scorecard Performance

4.4.1 Ranking Survival

After each passenger score was generated, passengers were aggregated into 20-point score bands and survival was calculated by band. The resulting survival curves show that our scorecards are able to rank the likelihood of survival with higher survivability as scores increase.

Table 5: Survival Rate by Score Band

	Train			Test			Validation		
Score	Died	Lived	Survival	Died	Lived	Survival	Died	Lived	Survival
385-440	251	21	8%	84	12	13%	79	21	21%
441-460	74	15	17%	15	4	21%	26	7	21%
461-480	48	13	21%	20	6	23%	20	8	29%
481-500	37	18	33%	10	7	41%	11	5	31%
501-520	53	48	48%	14	21	60%	21	13	38%
521-540	20	47	70%	11	27	71%	6	14	70%
541-560	4	35	90%	1	7	88%	0	17	100%
561-580	1	53	98%	0	11	100%	1	10	91%
581-633	1	35	97%	0	10	100%	1	15	94%

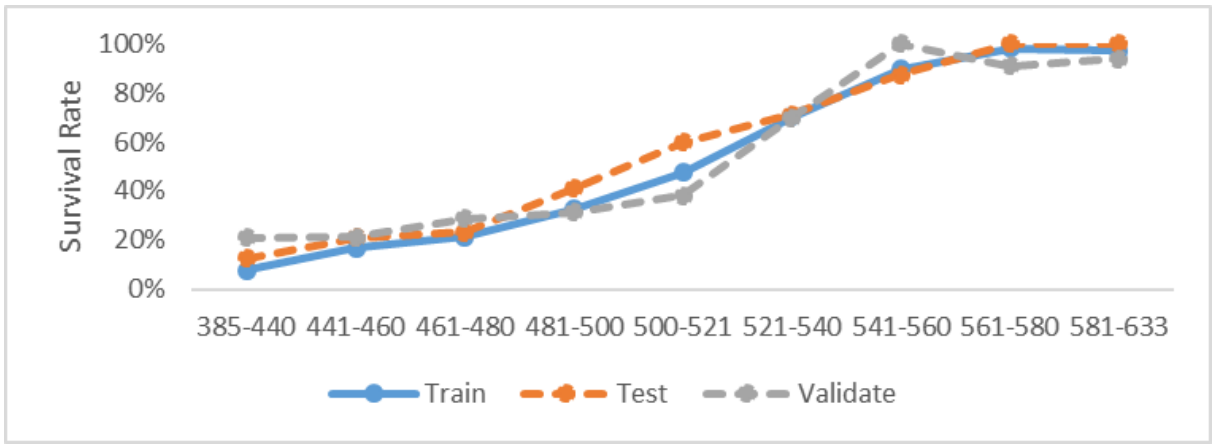

Figure 2: Survival Rate by Score Band

4.4.2 Scorecard Evaluation Metrics

The two-sample Kolmogorov-Smirnov (K-S) Statistic measures the maximum vertical separation of cumulative counts of two populations (MIT, 2006). We use the K-S critical value to evaluate whether to reject the null hypothesis that the cumulative distributions of the negatives (died) and positives (lived) are similar (Hartigan, 2019).

Equation 4: K-S Separation and Critical Value

$$KS\ Separation = \max|distribution\ negative - distribution\ positive|$$

$$KS\ Critical\ Value = c(a)\sqrt{\frac{negatives + positives}{negatives\ \times positives}}$$

$$if\ KS\ Separation > KS\ Critical\ Value: Reject\ H_0$$

To calculate the critical value, $c(a)$ is given at various levels of significance from the Kolmogorov distribution.

Table 6: K-S significance levels (Hartigan, 2019)

	Significance					
a level	<0.10	<0.05	<0.025	<0.01	<0.005	<0.001
$c(a)$	1.22	1.36	1.48	1.63	1.73	1.95

The Area Under the Curve (AUC) is a nonparametric two-sample test from the Receiver Operating Characteristic (ROC) Curve which is drawn from the cumulative population of negatives (dead) vs. cumulative population of positives (survived). The farther away the model ROC is from the random diagonal line, the closer AUC is to 1.00, and the more powerful the scorecard at predicting survival (Fawcett, 2006). The Cumulative Accuracy Profile (CAP) or Lorenz Curve, works similar to the ROC but is drawn from the total cumulative population vs. cumulative population of positives (survived) (Sobehard, Keenan, & Stein, 2000). In machine learning classification, a confusion matrix is used to gauge model performance by cross-tabulating observed positives and negatives against predicted positives and negatives, and measuring the total true positives and negatives detected against the total population.

Table 7: Model Evaluation (Skantzos & Castelein, 2016)

Model Predictive Power	Area Under the Curve (ROC/CAP)	Confusion Matrix Accuracy
Acceptable	>0.70	>0.60
Good	>0.80	>0.70
Very Good	>0.85	>0.85

Using these metrics, we find our scorecards have excellent predictive power in separating the dead and surviving populations (Test K-S 0.558, ROC 0.892, CAP 0.765; Validation K-S 0.461, AUC 0.842, CAP 0.731).

Table 8: Scorecard Evaluation Results

	Train	Test	Validation
K-S Statistic	0.603 (p < 0.001)	0.558 (p < 0.001)	0.461 (p < 0.001)
Area under the Curve (ROC)	0.914	0.892	0.842
Area under the Curve (CAP)	0.786	0.765	0.731

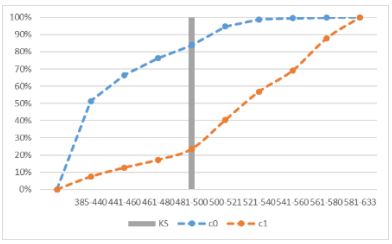

Figure 3: K-S (Train)

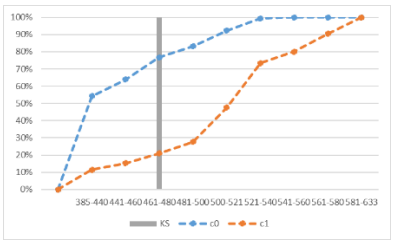

Figure 4: K-S (Test)

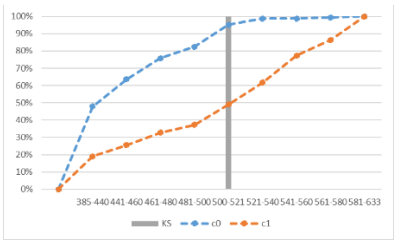

Figure 5: K-S (Validation)

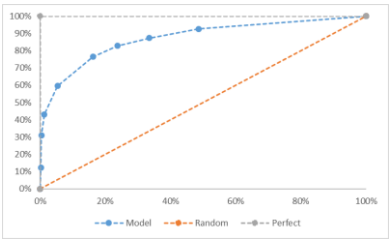

Figure 6: ROC (Train)

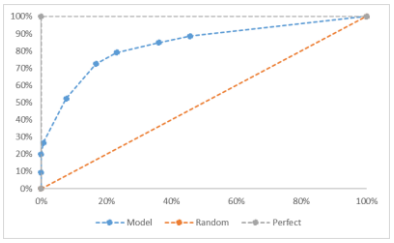

Figure 7: ROC (Test)

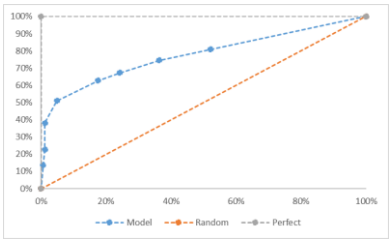

Figure 8: ROC (Validation)

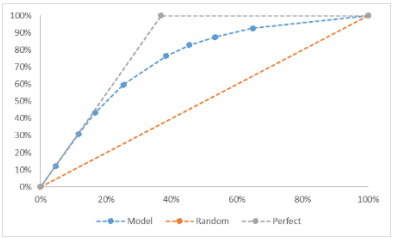

Figure 9: CAP (Train)

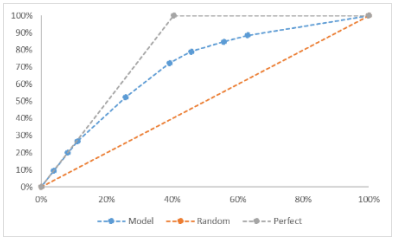

Figure 10: CAP (Test)

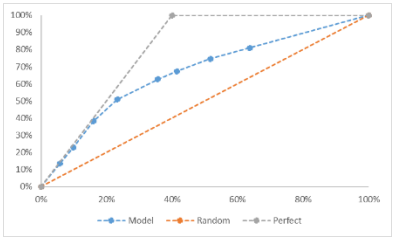

Figure 11: CAP (Validation)

4.4.3 Populating the Confusion Matrix

To calculate classification accuracy, a cut-off level must be selected by the user. Below the cut-off, passengers are predicted deceased (negative), and above the cut-off, passengers are predicted survived (positive). Score bands provide the user with various levels of accuracy, recall, precision and specificity, trading-off between true positives and true negatives. Scorecard accuracy will vary by cutoff, with the best overall outcome is a cut-off of 501 points, which has an accuracy of 84%, recall of 72%, precision of 75%, and specificity of 92%.

Table 9: Score Cut-off Confusion Matrix (Test)

Cutoff	Pred Neg PN	Pred Pos PP	True Neg TN	True Pos TP	Accuracy AR	Recall TP / P	Precision TP / PP	Specificity TN / N
385	0	260	84	105	73%	100%	40%	54%
441	96	164	99	93	74%	89%	57%	64%
461	115	145	119	89	80%	85%	61%	77%
481	141	119	129	83	82%	79%	70%	83%
501	158	102	143	76	84%	72%	75%	92%
521	193	67	154	55	80%	52%	82%	99%
541	231	29	155	28	70%	27%	97%	100%
561	239	21	155	21	68%	20%	100%	100%
581	250	10	155	10	63%	10%	100%	100%

4.5 Population Analysis

4.5.1 Risk by Attribute

Scorecards allow the disaggregation of the population by risk level through score distributions and average scores of a particular attribute. Using the scorecard, populations with a higher average score and populate the higher score bands have a higher likelihood of survival.

The average score of females is 19% higher than males and we observe female passengers populate the higher score bands compared to males. Passengers who embarked from Cherbourg have 7% higher average score than passengers from Queenstown and Southampton, and we observe the C passengers similarly populate higher score bands compared to Q and S passengers. Class 1 passengers have a 5% and 15% higher average score than Class 2 and Class 3 respectively and Class 1 also populates higher score bands than the other two. By disaggregating, the populations by score bands we are better able visualize the relative risks of each population attribute. We observe that the average scores per attribute line up well to the expected survival rate from the training dataset. We can also compare the respective average scores to the average score of the total population (479) to show relative survival risk.

Table 10: Score Distribution by Attribute (Test)

	Sex		Embarkation			Passenger Class			
Score Range	female	male	C	Q	S, blank	1	2	3	Total
385-440		96	7	8	81	1	8	87	96
441-460		19	4	3	12	3	7	9	19
461-480	2	24	6	1	19	13	6	7	26
481-500	4	13	4		13	9	4	4	17
501-520	27	8	6	7	22	8	4	23	35
521-540	30	8	11	4	23	9	18	11	38
541-560	7	1	2		6	5	1	2	8
561-580	11		3		8	10	1		11
581-633	10		6		4	10			10
Average Score	534	448	508	471	472	522	493	452	479
Survival Rate (Train)	72.5%	16.4%	54.6%	34.2%	31.7%	62.0%	43.0%	22.8%	36.8%

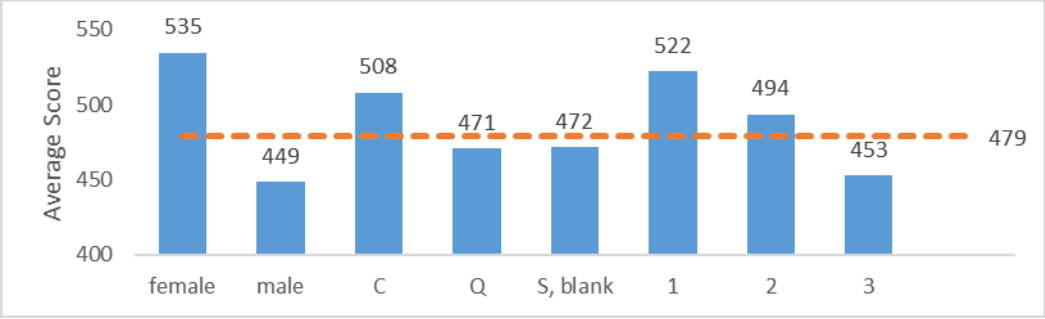

Figure 12: Average Scores Comparison (Test)

4.5.2 Population Shift, Population Stability

The calculation for Information Value can be used to gauge the shift or differences between populations, which can help explain the differences in relative risk. Population Stability Index (PSI) is a measure of differences in populations, and interpreted similar to IV – i.e. a PSI of 2.0 or less represents no change, while a value of 50.0 or more represents a significant shift. Any relative zero counts are ignored in calculating PSI.

Equation 5: Population Stability Index

$$PSI = \ln\left(\frac{Distribution\ Target}{Distribution\ Source}\right) \times (Distribution\ Target - Distribution\ Source) \times 100$$

Using PSI on the Passenger class feature, we see that Class 2 and 3 have an extremely high PSI relative to Class 1, which indicates that they are dissimilar populations from a risk perspective. The population shift shows where the differences are, generally Class 2 and 3 have more representation at the lowest score bands and lower representation at the higher score bands relative to Class 1. A chi-square test of independence for the Passenger class has an X2 value of 153.80 at 16 degrees of freedom with a $p < 0.0001$ confirming that the three classes are independent of each other.

Table 11: Population Stability Index - Passenger Class (Test)

	Population Distribution			Population Shift		Population Stability	
Score Range	Class 1	Class 2	Class 3	Class 1 to 2	Class 1 to 3	Class 1 to 2	Class 1 to 3
385-440	1%	16%	61%	15%	59%	35.76	221.00
441-460	4%	14%	6%	10%	2%	11.60	0.66
461-480	19%	12%	5%	-7%	-14%	3.06	19.37
481-500	13%	8%	3%	-5%	-10%	2.45	16.22
501-520	12%	8%	16%	-4%	4%	1.31	1.35
521-540	13%	37%	8%	23%	-6%	23.98	3.00
541-560	7%	2%	1%	-5%	-6%	6.80	9.88
561-580	15%	2%	0%	-13%	-15%	25.01	0
581-633	15%	0%	0%	-15%	-15%	0	0
					Total PSI	110.00	271.51

5. Conclusion and Recommendations for Future Study

In this paper we illustrated the development of a survival scorecard from the Titanic passenger dataset using statistical scoring, a method widely used in financial risk management. Although the Titanic dataset is a popular target for statistical and machine learning analysis, we have shown the benefits of statistical scoring methods not just in predicting survival risk but also providing a more intuitive analysis of the characteristics of a population with respect to the likelihood of a target event. The scorecard results confirm the findings of past studies on Titanic survival, with sex and social class (i.e. fare, cabin, passenger class) strongly influencing who survived the tragedy (Hall, 1986). While most existing machine learning approaches focus on maximizing classification accuracy (Farag & Hassan, 2018), scoring allows the user to dissect a population by risk level, and score cut-offs provide avenues to adjust classification accounting for tradeoffs between accuracy, recall, precision, and specificity.

The Titanic dataset's characteristics make it an ideal candidate for classification analysis (Harrell, 2001); however, it only represents a single snapshot of a risk event. In another context where populations may face risks that are time-dependent and evolve over a temporal dimension, scoring can also be used to measure how future risks compare and deviate from original conditions. The Titanic data also represent a 100% accept rate i.e. all passengers boarded the ship. Scorecard analysis can also involve a "reject inference" (Luca, 2018) – i.e. what if some passengers did not board the ship, would that change the risk characteristics of the event? Could we have improved survival by not accepting certain passengers from embarking? Although the ethical implications of such a study are dubious, there are other contexts where a similar problem formulation would be benign and applicable.

Acknowledgements

The author would like to thank Darwin, Sean, Simon, Kevin, Louise, Marga, Clark, Kitten, Aloy, Piluchi, Steph, Iris, Aki, Khai San, Albert, Amy, Meme, and all past collaborators in financial risk.

References

Cason, T., & Harrell, F. (2017). *Titanic*. From OpenML: https://www.openml.org/d/40945

Chaiyadecha, S. (2020). *A short story of Credit Scoring and Titanic dataset*. From medium.com: https://medium.com/analytics-vidhya/credit-scoring-model-9730d530f4ef

Dixon, R., & Griffiths, W. (2004). *Survival on the Titanic: Illustrating Wald and Lm Tests for Proportions and Logits*. doi:https://dx.doi.org/10.2139/ssrn.507182

Encylopedia Titanica. (1996-2021). From https://www.encyclopedia-titanica.org/.

Farag, N., & Hassan, G. (2018). Predicting the Survivors of the Titanic, Kaggle, Machine Learning From Disaster. *7th International Conference on Software and Information Engineering ICSIE '18*. doi:https://doi.org/10.1145/3220267.3220282

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. doi:https://doi.org/10.1016/j.patrec.2005.10.010

Hall, W. (1986). Social Class and Survival on the S.S. Titanic. *Social Science & Medicine, 22*(6), 687-690. doi:https://doi.org/10.1016/0277-9536(86)90041-9

Harrell, F. (2001). Logistic Model Case Study 2: Survival of Titanic Passengers. In *Regression Modeling Strategies. Springer Series in Statistics.* (pp. 299-330). Springer, New York, NY. doi:https://doi.org/10.1007/978-1-4757-3462-1_12

Hartigan. (2019). *Critical Values for the Two-sample Kolmogorov-Smirnov test (2-sided).* From sparky.rice.edu: https://sparky.rice.edu/astr360/kstest.pdf

Ibrahim, A., & Abdulaziz, R. (2020). Analysis of Titanic Disaster using Machine Learning Algorithms. *Engineering Letters, 28*(4), 1161-1167. From

https://www.researchgate.net/publication/353352089_Analysis_of_Titanic_Disaster_using_Machine_Learning_Algorithms

Kaggle. (2020). *The Complete Titanic Dataset*. From kaggle.com: https://www.kaggle.com/vinicius150987/titanic3

Lindemann, A., & Stolz, J. (2021). Teaching Mixed Methods: Using the Titanic Datasets to Teach Mixed Methods Data Analysis. *European Journal of Research Methods for the Behavioral and Social Sciences, 17*(3), 231-249. doi:https://doi.org/10.5964/meth.4241

Luca, S. (2018). *Evaluation of Different Approaches to Reject Inference: a case study in Credit Risk*. From SAS.com: https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2731-2018.pdf

MIT. (2006). *Section 13 - Kolmogorov-Smirnov Test*. From ocw.mit.edu: https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture14.pdf

Nair, P. (2017). Analyzing Titanic Disaster using Machine Learning Algorithms. *International Journal of Trend in Scientific Research and Development, 2*(1), 410-416. doi:https://doi.org/10.31142/ijtsrd7003

Nikitina, N., & Zamnius, A. (2019). *How to Survive the Titanic?* doi:https://dx.doi.org/10.2139/ssrn.3319888

Ranganathan, P., Pramesh, C., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. *Perspectives in Clinical Research, 8*(3), 148-151. doi:10.4103/picr.PICR_87_17

Siddiqi, N. (2017). *Intelligent Credit Scoring*. John Wiley & Sons, Inc. .

Skantzos, N., & Castelein, N. (2016). *Credit Scoring - Case study in data analytics*. From Deloitte.com: https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Financial-Services/gx-be-aers-fsi-credit-scoring.pdf

Sobehard, J., Keenan, S., & Stein, R. (2000). *Validation methodologies for default risk models*. Moody's Investors Service. From http://www.rogermstein.com/wp-content/uploads/SobehartKeenanStein2000.pdf

Tables, Figures, and Equations