

SAS Studio Capstone

by

Joshua Troup

A capstone Report submitted in partial fulfillment of the requirements for the degree

Master of Science in Applied Data Analytics

Department of Computer Information Science

Clarion University of Pennsylvania

Clarion, PA

March 2018

Capstone Supervisor:

Dr. Jayakumar V Annadatha, Associate Professor

Contents

1. Abstract/Introduction	3
2. Project and Data Selection	5
3. Exploring the data	7
4. Modify Data	9
4.1 Working with Missing values	10
4.2 Data Partitioning	12
4.3 Managing Outliers	13
4.4 Transformation of Variables	13
5. Development of Model(s)	15
5.1 Model 1 (LR All variables)	15
5.2 Model 2 (LR Select variables)	16
6. Analytical Findings/Insights.....	17
7. Conclusions.....	22
8. References	23

1. Abstract/Introduction

Big data can be classified as a large set of complex data that is either structured, semi-structured, or non-structured that is not easily processed by the human eye or traditional applications. Technology advancements have changed the ways we do things in life and the workplace including ways of working with the astronomical amount of data being generated by countless devices across the world. Big data analytics consisting of analyzing and examining vast amounts of data to discover hidden insights, patterns, correlations, among other valuable information a company or organization may benefit from. There are four main types of big data analytics methods that are used most frequently in businesses today. Prescriptive analytics determines what actions should be taken involving rules and recommendations. Diagnostic analytics uses past performance to determine what happened and why usually in a dashboard format. Descriptive analytics uses real time incoming data and data mining is typically performed. Predictive analytics uses likely scenarios of what might happen and delivered as a predictive forecast. These are the four most used types of analytics used today however for my report I will focus mainly on a major crisis that happened over 105 years ago using predictive analysis.

SAS Studio is a powerful statistical cloud based software providing an assistive framework for programmers and novice users allowing access anywhere an internet connection is present. SAS offers autocomplete, syntax help, color coding, keyboard shortcuts and other features to make programming easier for the programmer. SAS predictive analytics allow the user to modify the data with ease by structuring the data set, transpose them, and aggregate values into representative modeling tables. The ability to quickly turn models into visualizations

from multiple angles throughout many of the functions built into SAS. “Powerful statistical analysis is the foundation of SAS solutions. Use it to perform everything from simple descriptive statistics to complex Bayesian analyses, including variance analysis, categorical data analysis, survival analysis, predictive modeling, experimental design, time series analysis, clustering and survey data analysis.” (SAS Predictive Analytics, 2018)

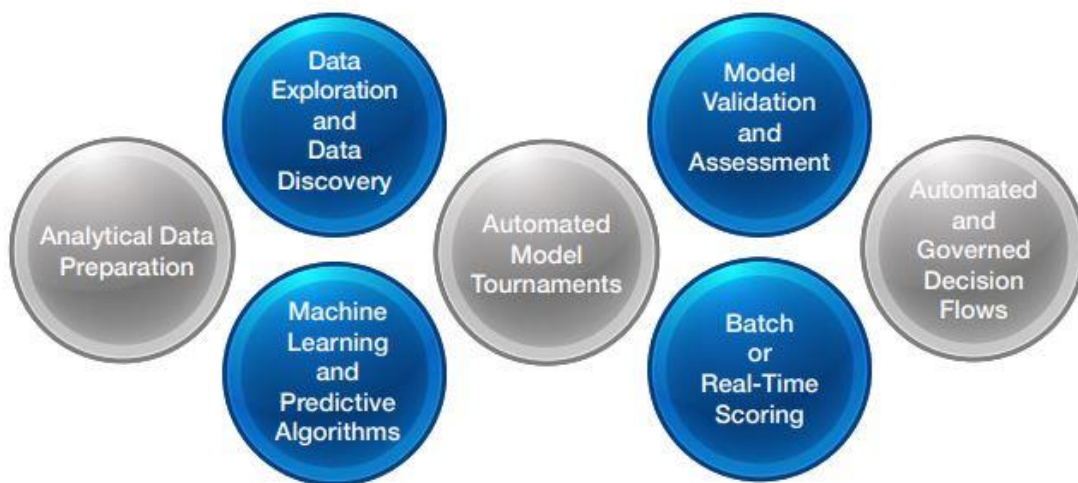


Figure 1: SAS predictive analytics key functional areas.

Fig 1

Fig 1 shows the multiple functional areas within SAS predictive analytics. Quantifying uncertainty to drive decisions with the correct method and measure the predictive strength of models using test and validation data sets to gain the most accurate results. The machine learning features built in allow for the user to reuse models, retrain predictive models with new data, and automate decision flows.

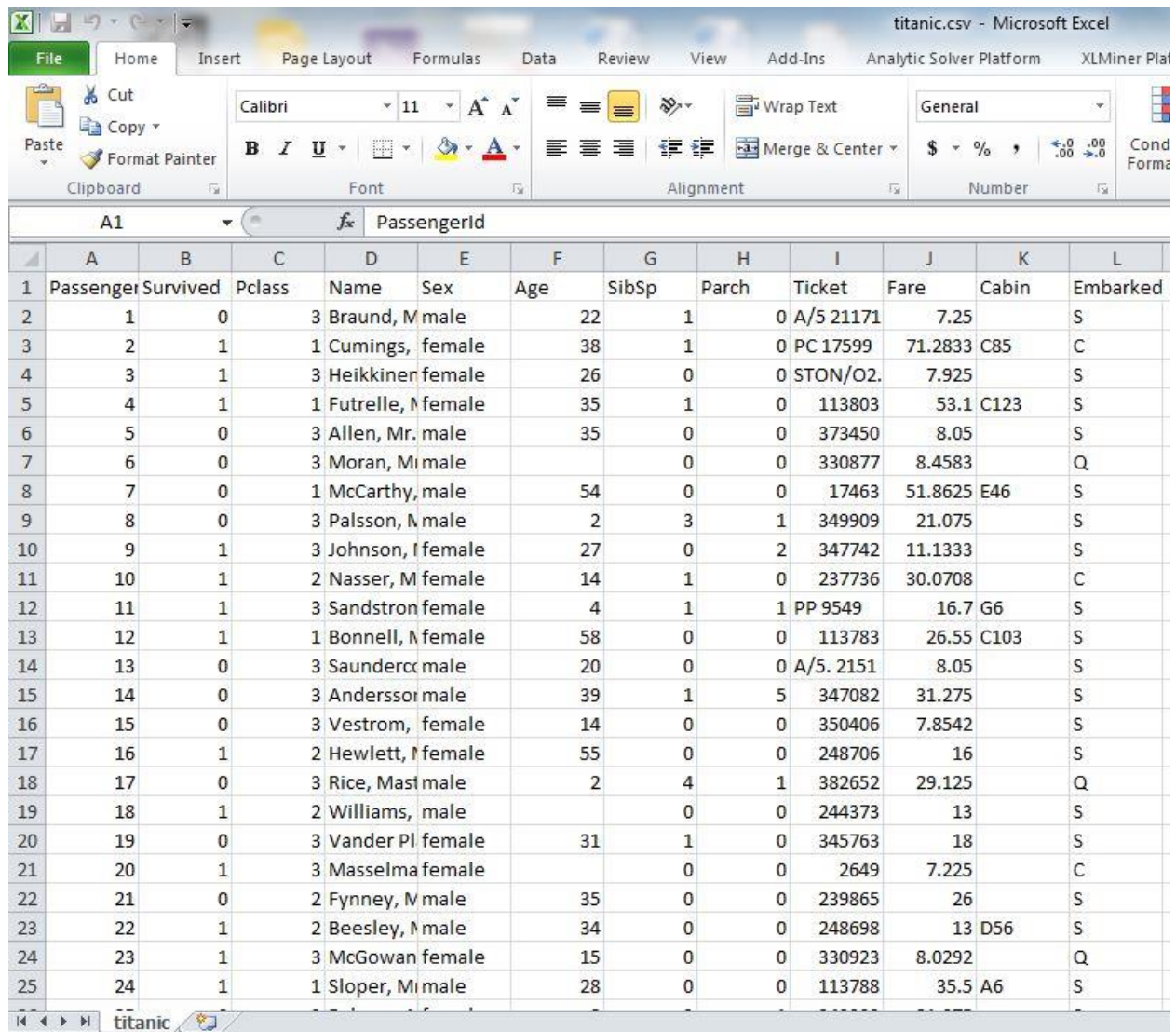
2. Project and Data Selection

The Royal Mail Ship (RMS) Titanic goes down in history as one of the most infamous shipwrecks which sank after hitting an iceberg to the bottom of the ocean on April 15, 1912. 1,502 people were killed out of the 2,224 passengers and crewmen. One of the biggest reasons led to believe was the insufficient amount of the lifeboats to support all 2,224 passengers was one of the largest factors with women, children, and upper-class passengers more likely to board first increasing survival rate. The analytical objective is to apply predictive analytics technique to this classic case of Titanic survivors. Can I predict passenger survival using a supervised machine learning technique? To meet my goal, I will be using Logistic Regression algorithm. Passengers will be classified using the binary labels: 1 for those who survived and 0 for those who did not. Fig 2 below includes a chart with all the variables in the data set along with definitions and keys.

Variable	Definition	Key
Survival	Survival	0 = No, 1 = Yes
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	Sex	
Age	Age in years	
Sibsp	# of siblings / spouses aboard the Titanic	
Parch	# of parents / children aboard the Titanic	
Ticket	Ticket number	
Fare	Passenger fare	
Cabin	Cabin number	
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Fig 2

3. Exploring the data



	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, female	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, female	female	26	0	0	STON/O2.	7.925		S
5	4	1	1	Futelle, M	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, M	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, male	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, M	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom, female	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, M	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saunders, male	male	20	0	0	A/5. 2151	8.05		S
15	14	0	3	Andersson, male	male	39	1	5	347082	31.275		S
16	15	0	3	Vestrom, female	female	14	0	0	350406	7.8542		S
17	16	1	2	Hewlett, M	female	55	0	0	248706	16		S
18	17	0	3	Rice, Master	male	2	4	1	382652	29.125		Q
19	18	1	2	Williams, male	male		0	0	244373	13		S
20	19	0	3	Vander Planck, female	female	31	1	0	345763	18		S
21	20	1	3	Masella, female	female		0	0	2649	7.225		C
22	21	0	2	Fynney, M	male	35	0	0	239865	26		S
23	22	1	2	Beesley, M	male	34	0	0	248698	13	D56	S
24	23	1	3	McGowan, female	female	15	0	0	330923	8.0292		Q
25	24	1	1	Sloper, M	male	28	0	0	113788	35.5	A6	S

Fig 3

The raw unprocessed titanic.csv file is displayed above in Fig 3 in Excel. The dataset contains missing values that will need to be addressed further in this report.

```

1 proc import datafile="/home/jltroup10/sas files example/SAS CAPSTONE/titanic.csv" out = train
2 dbms = csv
3 replace;
4 guessingrows=2000;
5 run;

```

Fig 4 Import dataset syntax

Table: WORK.TRAIN View: Column names Filter: (none)

Columns: Select all, PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch

Total rows: 891 Total columns: 12 Rows 1-100

	PassengerId	Survived	Pclass	Name
1	1	0	3	Braund, M
2	2	1	1	Cumings,
3	3	1	3	Heikkinen
4	4	1	1	Futrelle, M
5	5	0	3	Allen, Mr.
6	6	0	3	Moran, Mi
7	7	0	1	McCarthy,
8	8	0	3	Palsson, M
9	9	1	3	Johnson,
10	10	1	2	Nasser, M
11	11	1	3	Sandstrom
12	12	1	1	Bonnell, M
13	13	0	3	Saunders,
14	14	0	3	Andersson
15	15	0	3	Vestrom, M

Messages: 2 User: jltroup10

Fig 4.1

Fig 4 and 4.1 displays the syntax and output of the csv dataset in SAS Studio for exploration.

4. Modify Data

```

7 DATA train;
8   SET train;
9   IF Sex = "male" THEN Sex_numeric = 1;
10  IF Sex = "female" THEN Sex_numeric = 0;
11  IF Embarked = "C" THEN Embarked_C = 1; ELSE Embarked_C = 0;
12  IF Embarked = "Q" THEN Embarked_Q = 1; ELSE Embarked_Q = 0;
13  /*IF Embarked = "S" THEN Embarked_S = 1; ELSE Embarked_S = 0;*/
14 RUN;
15 PROC PRINT DATA=train; RUN;

```

Fig 5 Modifying data with IF, THEN syntax

Obs	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Sex_numeric	Embarked_C	Embarked_Q
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S	1	0	0
2	2	1	1	Cummings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C	0	1	0
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S	0	0	0
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S	0	0	0
5	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S	1	0	0
6	6	0	3	Moran, Mr. James	male	.	0	0	330877	8.4583		Q	1	0	1
7	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S	1	0	0
8	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S	1	0	0
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S	0	0	0
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C	0	1	0
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S	0	0	0

Fig 5.1

Fig 5 and 5.1 display the syntax and output of modifying the data. Sex and Embarked were created into dummy variables since they are not numerical but considered relevant variables.

4.1 Working with missing values

```
17 PROC MEANS DATA=train N NMISS MIN P1 P5 P50 P95 P99 MAX MIN;  
18 VAR Pclass Age SibSp Parch Fare;  
19 RUN;
```

Fig 6 Missing value in variables syntax

The MEANS Procedure									
Variable	N	N Miss	Minimum	1st Pctl	5th Pctl	50th Pctl	95th Pctl	99th Pctl	Maximum
Pclass	891	0	1.0000000	1.0000000	1.0000000	3.0000000	3.0000000	3.0000000	3.0000000
Age	714	177	0.4200000	1.0000000	4.0000000	28.0000000	56.0000000	66.0000000	80.0000000
SibSp	891	0	0	0	0	0	3.0000000	5.0000000	8.0000000
Parch	891	0	0	0	0	0	2.0000000	4.0000000	6.0000000
Fare	891	0	0	0	7.2250000	14.4542000	113.2750000	262.3750000	512.3292000

Fig 6.1

Fig 6 and 6.1 display the missing values in the numerical variables. Age has 177 missing variables with all other variables showing no null values.

```
21 PROC FREQ DATA=train; TABLES Sex Cabin Embarked;  
22 RUN;
```

Fig 7 Missing value non-numerical syntax

Embarked	Frequency	Percent	Cumulative Frequency	Cumulative Percent
C	168	18.90	168	18.90
Q	77	8.66	245	27.56
S	644	72.44	889	100.00
Frequency Missing = 2				

Fig 7.1

Fig 7 and 7.1 show the missing values in discrete non-numerical variables. Two missing values shown in the Embarked variable.

```

24 DATA train;
25     SET train;
26     IF Age = . THEN Age = 32.75;
27     IF Embarked = . THEN Embarked = "S";
28 RUN;
29 PROC PRINT DATA=train; RUN;

```

Fig 8 Replacing missing values with median syntax

bs	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Sex_numeric	Embarked_C	Embarked_Q
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S	1	0	0
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	S	0	1	0
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S	0	0	0
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S	0	0	0
5	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S	1	0	0
6	6	0	3	Moran, Mr. James	male	32.75	0	0	330877	8.4583		S	1	0	1
7	7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S	1	0	0
8	8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S	1	0	0
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S	0	0	0
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		S	0	1	0
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S	0	0	0

Fig 8.1

Fig 8 and 8.1 displays the missing numerical values in age have been replaced with a median of 32.75. Embarked missing values have been replaced with S.

4.2 Data Partitioning

```
31 PROC SURVEYSELECT DATA=train outall OUT=all METHOD=srs SAMPRATE=0.3;  
32 RUN;  
33 PROC FREQ DATA=all; TABLES selected; RUN;  
34 DATA dev;  
35     SET all;  
36     IF selected = 1 THEN DELETE; /*1 is test data*/  
37 RUN;  
38 DATA test;  
39     SET all;  
40     IF selected = 0 THEN DELETE; /*0 is development data*/  
41 RUN;  
42 PROC SQL;  
43     SELECT COUNT(*) FROM train;  
44     SELECT COUNT(*) FROM dev;  
45     SELECT COUNT(*) FROM test;  
46 RUN;
```

Fig 9 Data partitioning syntax

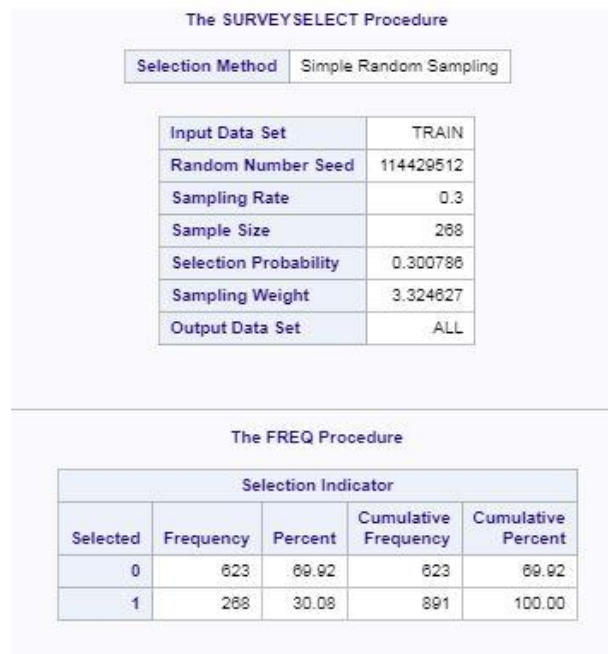


Fig 9.1

Fig 9 and 9.1 display the data partitioning 70% into training/development and 30% into test set.

4.3 Managing Outliers

```
48 PROC MEANS DATA=train MIN P1 P5 P95 P99 MAX ;  
49 VAR Age Fare;  
50 RUN;
```

Fig 10 Percentiles syntax

The MEANS Procedure						
Variable	Minimum	1st Pctl	5th Pctl	95th Pctl	99th Pctl	Maximum
Age	0.4200000	1.0000000	6.0000000	54.0000000	65.0000000	80.0000000
Fare	0	0	7.2250000	113.2750000	262.3750000	512.3292000

Fig 10.1

Fig 10 and 10.1 displays the percentiles of the variables age and fare. The means procedure shows there are no outliers so no capping or flooring is needed.

4.4 Transformation of Variables

```
52 DATA train;  
53 SET train;  
54 IF Pclass = 1 THEN Pclass_1 = 1; ELSE Pclass_1=0;  
55 IF Pclass = 2 THEN Pclass_2 = 1; ELSE Pclass_2=0;  
56 IF Pclass = 3 THEN Pclass_3 = 1; ELSE Pclass_3=0;  
57  
58 IF Age<18 THEN Child = 1; ELSE Child=0;  
59 IF 18<=Age<50 THEN Adult = 1; ELSE Adult=0;  
60 IF 50<=Age THEN Senior = 1; ELSE Senior=0;  
61 RUN;
```

Fig 11 IF THEN transformation syntax

Pclass_3	Child	Adult	Senior
1	0	1	0
0	0	1	0
1	0	1	0
0	0	1	0
1	0	1	0
1	0	1	0
0	0	0	1
1	1	0	0
1	0	1	0
0	1	0	0
1	1	0	0
0	0	0	1
1	0	1	0
1	0	1	0

Fig 11.1

Figure 11 and 11.1 show transformation of Pclass and Age variable.

Dependent - Survived

Independent – Adult, Age, Child, Embarked_C, Embarked_Q, Fare, Parch, Pclass_1, Pclass_2, Sex_numeric, SibSp

(Note: Senior and Pclass_2 dummy variables not required as other relevant dummy variables are present.)

5. Development of Models

The development of models will be an experimentation stage of the analysis. Multiple models will be developed experimenting with a variety of variables to determine the best fitting model to validate and test in later stages that will deliver the great amount of accuracy in predictive analytics.

5.1 Model 1 – Logistic Regression with all variables

```
63 PROC LOGISTIC DATA=dev OUTMODEL=model_1 OUTEST = estimates_1;  
64     MODEL Survived =      Adult  
65                          Age  
66                          Child  
67                          Embarked_C  
68                          Embarked_Q  
69                          Fare  
70                          Parch  
71                          Pclass_1  
72                          Pclass_2  
73                          Sex_numeric  
74                          SibSp ;  
75 RUN;  
76 PROC PRINT DATA=model_1; RUN;  
77 PROC PRINT DATA=estimates_1; RUN;
```

Fig 12 Logistic Regression all variables syntax

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.4159	0.9582	0.1884	0.6642
Adult	1	-0.5047	0.5509	0.8394	0.3596
Age	1	0.0149	0.0160	0.8715	0.3505
Child	1	-1.6163	0.8895	3.3014	0.0692
Embarked_C	1	-0.3733	0.2823	1.7485	0.1861
Embarked_Q	1	-0.3812	0.4031	0.8941	0.3444
Fare	1	-0.00119	0.00311	0.1472	0.7012
Parch	1	0.0745	0.1651	0.2036	0.6518
Pclass_1	1	-1.9912	0.3598	30.6257	<.0001
Pclass_2	1	-1.2534	0.2893	18.7655	<.0001
Sex_numeric	1	2.7898	0.2506	123.9009	<.0001
SibSp	1	0.4239	0.1371	9.5649	0.0020

Fig 12.1

Fig 12 and 12.1 show model one using logistic regression and all variables indicated in variable glossary. Using a 0.05 threshold, only Pclass_1, Pclass_2, Sex_numeric, and SibSp are relevant and less than threshold according to Pr > ChiSq in model 1. Model two includes the variable Child since it shows to be slightly above the 0.05 threshold.

5.2 Model 2 – Logistic Regression with select variables

```

118 PROC LOGISTIC DATA=dev OUTMODEL=model_2 OUTEST = estimates_2;
119     MODEL Survived =
120         Child
121         Pclass_1
122         Pclass_2
123         Sex_numeric
124         SibSp;
125 RUN;
126 TITLE "Estimates for second run";
127 PROC PRINT DATA=estimates_2; RUN;

```

Fig 13 Logistic Regression selected variables syntax

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.6066	0.2218	7.4821	0.0062
Child	1	-1.4440	0.3484	17.1763	<.0001
Pclass_1	1	-1.8363	0.2649	48.0674	<.0001
Pclass_2	1	-1.1047	0.2747	16.1745	<.0001
Sex_numeric	1	2.8373	0.2367	143.6645	<.0001
SibSp	1	0.4417	0.1305	11.4580	0.0007

Fig 13.1

Model two shown in Fig 13 and 13.1 displays the five significant variables from model one all well below the threshold of 0.05.

Coefficient and Intercept estimate equation as the following:

$\log(p/(1-p)) = -0.60656 + (-1.44404 * \text{Child}) + (-1.83631 * \text{Pclass}_1) + (-1.10466 * \text{Pclass}_2) + (2.83730 * \text{Sex_numeric}) + (0.44172 * \text{SibSp})$ (Fig 14 shown below)

Estimates for second run												
Obs	_LINK_	_TYPE_	_STATUS_	_NAME_	Intercept	Child	Pclass_1	Pclass_2	Sex_numeric	SibSp	_LNLIKE_	_ESTTYPE_
1	LOGIT	PARMS	0 Converged	Survived	-0.60666	-1.44404	-1.83631	-1.10466	2.83730	0.44172	-275.344	MLE

Fig 14

6. Analytical Findings/Insights

```

136 PROC SCORE DATA=test
137 SCORE=estimates_2
138 OUT=scored_output
139 TYPE=PARMS;
140 VAR Child
141     Pclass_1
142     Pclass_2
143     Sex_numeric
144     SibSp ;
145 RUN;
146 TITLE "log odds in test score";
147 PROC PRINT DATA=scored_output(OBS=10); RUN;
148
149 DATA scored_final;
150     SET scored_output;
151     Pred_Prob=EXP(Survived2)/(1+EXP(Survived2));
152 RUN;
153 TITLE "Final probabilities";
154 PROC PRINT DATA=scored_final(OBS=10); RUN;
--

```

Fig 15 Test model two log odds & final probabilities output syntax

Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Sex_numeric	Embarked_C	Embarked_Q	Pclass_1	Pclass_2	Pclass_3	Child	Adult	Senior	Survived2
3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S	1	0	0	0	0	1	0	1	0	2.91547
3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S	1	0	0	0	0	1	1	0	0	2.23550
2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		S	0	1	0	0	1	0	1	0	0	-2.76654
3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S	0	0	0	0	0	1	1	0	0	-1.55978
3	Saunderscook, Mr. William Henry	male	20	0	0	A/5 2151	8.05		S	1	0	0	0	0	1	0	1	0	2.41584
2	Williams, Mr. Charles Eugene	male	32.75	0	0	244373	13		S	1	0	0	0	1	0	0	1	0	1.20909
3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31	1	0	345763	18		S	0	0	0	0	0	1	0	1	0	0.11945
2	Beesley, Mr. Lawrence	male	34	0	0	249008	13	D56	S	1	0	0	0	1	0	0	1	0	1.20909
3	Todoroff, Mr. Lailo	male	32.75	0	0	349216	7.8958		S	1	0	0	0	0	1	0	1	0	2.41584
1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	32.75	1	0	PC 17569	146.5208	B78	S	0	1	0	1	0	0	0	1	0	-2.00356

Fig 15.1

PassengerId	Survived	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Sex_numeric	Embarked_C	Embarked_Q	Pclass_1	Pclass_2	Pclass_3	Child	Adult	Senior	Survived2	Pred_Prob
1. Mr. John Harris	0	male	22	1	0	A/5 21171	7.25		S	1	0	0	0	0	1	0	1	0	2.91547	0.94861
2. Mrs. Gosta Larson	0	male	2	3	1	349009	21.075		S	1	0	0	0	0	1	1	0	0	2.23550	0.90339
3. Mrs. Adele Smith	0	female	14	1	0	237736	30.0708		S	0	1	0	0	1	0	1	0	0	-2.76654	0.05916
4. Mrs. Rutger Jueter	0	female	4	1	1	PP 9549	16.7	G6	S	0	0	0	0	0	1	1	0	0	-1.55978	0.17388
5. Mr. William Brock	0	male	20	0	0	A/5 2151	8.05		S	1	0	0	0	0	1	0	1	0	2.41584	0.91803
6. Mr. Les Smith	0	male	32.75	0	0	244373	13		S	1	0	0	0	1	0	0	1	0	1.20909	0.77014
7. Mr. Julius Plankie	0	female	31	1	0	345763	18		S	0	0	0	0	0	1	0	1	0	0.11945	0.52993
8. Mr. Jemmoentele	0	male	34	0	0	248998	13	D56	S	1	0	0	0	1	0	0	1	0	1.20909	0.77014
9. Mr. Vance	0	male	32.75	0	0	349218	7.8958		S	1	0	0	0	0	1	0	1	0	2.41584	0.91803
10. Mrs. Justus	0	female	32.75	1	0	PC 17569	145.5208	B78	S	0	1	0	1	0	0	0	1	0	-2.00356	0.11883

Fig 15.2

Fig 15 is SAS syntax to test model two with the five selected variables giving log odds and final probabilities output. Fig 15.1-15.2 are the log odds in test score from model two. We now have the survival probability (Pred_Prob) variable calculated to see how well our model predicts passengers who survived or the ones who did not.

```

27 PROC PRINT DATA=scored_final;
28 VAR PassengerId Survived Pred_Prob;
29 RUN;

```

Fig 16 Table creation with PassengerId, Survived, Pred_Prob

Fig 16 shows the syntax used to create a table displaying the PassengerId, Survived, and Pred_Prob that will be used to develop a correlation matrix based on variety different thresholds.

Obs	PassengerId	Survived	Pred_Prob
1	1	0	0.94861
2	8	0	0.90339
3	10	1	0.05916
4	11	1	0.17368
5	13	0	0.91803
6	18	1	0.77014
7	19	0	0.52983
8	22	1	0.77014
9	30	0	0.91803
10	32	1	0.11883
11	33	1	0.40608
12	36	0	0.68836
13	37	1	0.91803
14	40	1	0.17368
15	42	0	0.25213
16	44	1	0.05916
17	48	1	0.40608
18	50	0	0.52983
19	54	1	0.25213
20	61	0	0.91803
21	65	0	0.57289
22	75	1	0.91803

Fig 16.1

Fig 16.1 displays the three variable chart showcasing a screenshot of 22 out of 268 observations.

Predicted with 0.1 threshold			Original Response Rate	40%
Count of PassengerId	Column Labels			
Row Labels	0	1	Grand Total	
0	161	161	0's Predicted Correctly	0.00%
1	22	85	1's Predicted Correctly	79.44%
Grand Total	22	246		

Predicted with 0.2 threshold				
Count of PassengerId	Column Labels			
Row Labels	0	1	Grand Total	
0	3	158	0's Predicted Correctly	1.86%
1	51	56	1's Predicted Correctly	52.34%
Grand Total	54	214		

Predicted with 0.3 threshold				
Count of PassengerId	Column Labels			
Row Labels	0	1	Grand Total	
0	4	157	0's Predicted Correctly	2.48%
1	58	49	1's Predicted Correctly	45.79%
Grand Total	62	206		

Predicted with 0.4 threshold				
Count of PassengerId	Column Labels			
Row Labels	0	1	Grand Total	
0	20	141	0's Predicted Correctly	12.42%
1	71	36	1's Predicted Correctly	33.64%
Grand Total	91	177		

Predicted with 0.5 threshold				
Count of PassengerId	Column Labels			
Row Labels	0	1	Grand Total	
0	31	130	0's Predicted Correctly	19.25%
1	77	30	1's Predicted Correctly	28.04%
Grand Total	108	160		

Predicted with 0.6 threshold				
Count of PassengerId	Column Labels			
Row Labels	0	1	Grand Total	
0	48	113	0's Predicted Correctly	29.81%
1	89	18	1's Predicted Correctly	16.82%
Grand Total	137	131		

Predicted with 0.7 threshold				
Count of PassengerId	Column Labels			
Row Labels	0	1	Grand Total	
0	56	105	0's Predicted Correctly	34.78%
1	94	13	1's Predicted Correctly	12.15%
Grand Total	150	118		

Predicted with 0.8 threshold				
Count of PassengerId	Column Labels			
Row Labels	0	1	Grand Total	
0	79	82	0's Predicted Correctly	49.07%
1	96	11	1's Predicted Correctly	10.28%
Grand Total	175	93		

Predicted with 0.9 threshold				
Count of PassengerId	Column Labels			
Row Labels	0	1	Grand Total	
0	87	74	0's Predicted Correctly	54.04%
1	96	11	1's Predicted Correctly	10.28%
Grand Total	183	85		

Fig 16.2

7. Conclusions

The logistic models discovered sets of variables bearing statistically significant impacts on the likelihood of outcome of passengers on the Titanic shipwreck. The first model was developed with all variables to pinpoint significance in variable selection. The relative importance of predictors were decided by comparing the Test Chi-Square DF Pr > ChiSq value below a Threshold of 0.05. Model two includes all significant predictor variables with Test Chi-Square DF Pr > ChiSq below 0.05. Fig 16.2 displays the correlation matrix for all 268 observations with an original response rate of 40%. 0.1 Threshold indicates a strong correlation with (1) survived at a 79.44% and 0% of (0) non-survival. 0.2 Threshold indicates a fairly strong correlation with (1) survived at a 52.34% and a 1.86% of (0) non survival. 0.3 Threshold indicates a mildly strong correlation with (1) survived at a 45.79% and a 2.48% of (0) non survival. The pattern as the threshold increases continues to decrease in correlation (1) survived and increase (0) non survival. The five variables chosen with the smaller the threshold seem to be a good indicator of which passengers survived. Passenger Class (Pclass_1 & Pclass_2), Child, and Sex (Sex_numeric) are all strong correlating variables that indicate a strong case for survival. The saying “Women and children first” does seem to hold true based on the insights discovered in the data set. Siblings and Spouses (SibSp) variable is significant as well with many siblings (particularly female) and spouses being overly crowded in life boats. I predict not all siblings were female as the data does not portray the cowards fleeing to lifeboats with the women and children.

8. References

Declues, J. (n.d.). Four Types of Big Data Analytics and Examples of Their Use. Retrieved February 28, 2018, from <http://www.ingrammicroadvisor.com/data-center/four-types-of-big-data-analytics-and-examples-of-their-use>

(n.d.). Retrieved February 26, 2018, from <https://www.kaggle.com/c/titanic>

<https://support.sas.com/en/support-home.html>

Sas.com. (2018). [online] Available at:

https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-studio-107188.pdf [Accessed 26 Feb. 2018].

Sas.com. (2018). [online] Available at:

https://www.sas.com/content/dam/SAS/en_us/doc/overviewbrochure/sas-predictive-analytics-suite-103121.pdf [Accessed 26 Feb. 2018].