Titanic Challenge

Junchen Pan
Oct 2018

Background:

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, we ask you to complete the analysis of what sorts of people were likely to survive. In particular, we ask you to apply the tools of machine learning to predict which passengers survived the tragedy.

Dataset Statistics & Notes:

- Training set size : 890 rows
- Test set : 418 rows
- Total size : 1308 rows
- 11 features(Independent variables)
- 1 result(dependent variables)
- Note: Missing 218 entries for Age, 100 entries for Cabin and several entries for Embarked

Glimpse of Data:

Passenge	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heat	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leo	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Eli	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Ade	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerit	female	4	1	1	PP 9549	16.7	G6	S
12	1	1	Bonnell, Miss. Elizabeth	female	58	0	0	113783	26.55	C103	S
13	0	3	Saundercock, Mr. William H	male	20	0	0	A/5. 2151	8.05		S
14	0	3	Andersson, Mr. Anders Joha	male	39	1	5	347082	31.275		S
15	0	3	Vestrom, Miss. Hulda Amal	female	14	0	0	350406	7.8542		S
16	1	2	Hewlett, Mrs. (Mary D King	female	55	0	0	248706	16		S
17	0	3	Rice, Master. Eugene	male	2	4	1	382652	29.125		Q
18	1	2	Williams, Mr. Charles Euge	male		0	0	244373	13		S

Variable Meaning:

Variable	Definition
survival	Survival
pclass	Ticket class
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

Plan of Attack:

- I have decided to predict the Ages first using other relevant data and then combine them back to predict the survival rate
- I categorize Fare and Age continuous variables, and name text into one category variable
- I combining sibsp and parch into one new variable called Family, as family can show the relation for age and survival better instead of these two separately
- I dropped Embark, ticket, cabin, sibsp, parch
- I used RandomForest to determine the missing ages
- I used 3 layers NN to do the classification

Data Tranformation:

Raw Data:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

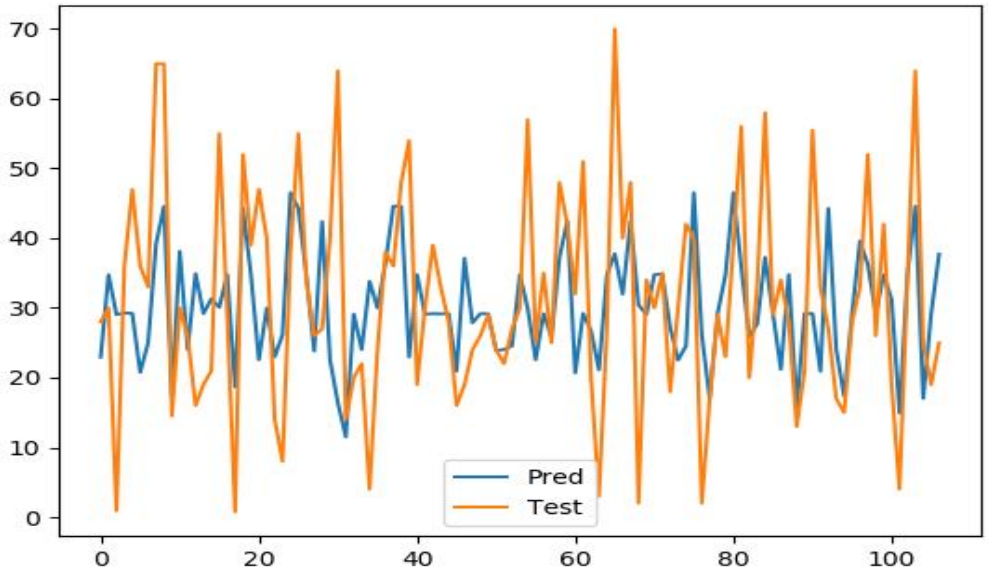
After Transformation:

	Age	Sex	Survived	Master	Miss	Mr	Mrs	Family	Bins_Fare	class1	Class2
0	4.0	1.0	0.0	0.0	0.0	1.0	0.0	2.0	1.0	0.0	0.0
1	5.0	0.0	1.0	0.0	0.0	0.0	1.0	2.0	4.0	1.0	0.0
2	4.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	2.0	0.0	0.0
3	5.0	0.0	1.0	0.0	0.0	0.0	1.0	2.0	4.0	1.0	0.0
4	5.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	2.0	0.0	0.0
5	4.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	2.0	0.0	0.0
6	6.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	4.0	1.0	0.0
7	1.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0
8	4.0	0.0	1.0	0.0	0.0	0.0	1.0	3.0	2.0	0.0	0.0
9	3.0	0.0	1.0	0.0	0.0	0.0	1.0	2.0	3.0	0.0	1.0

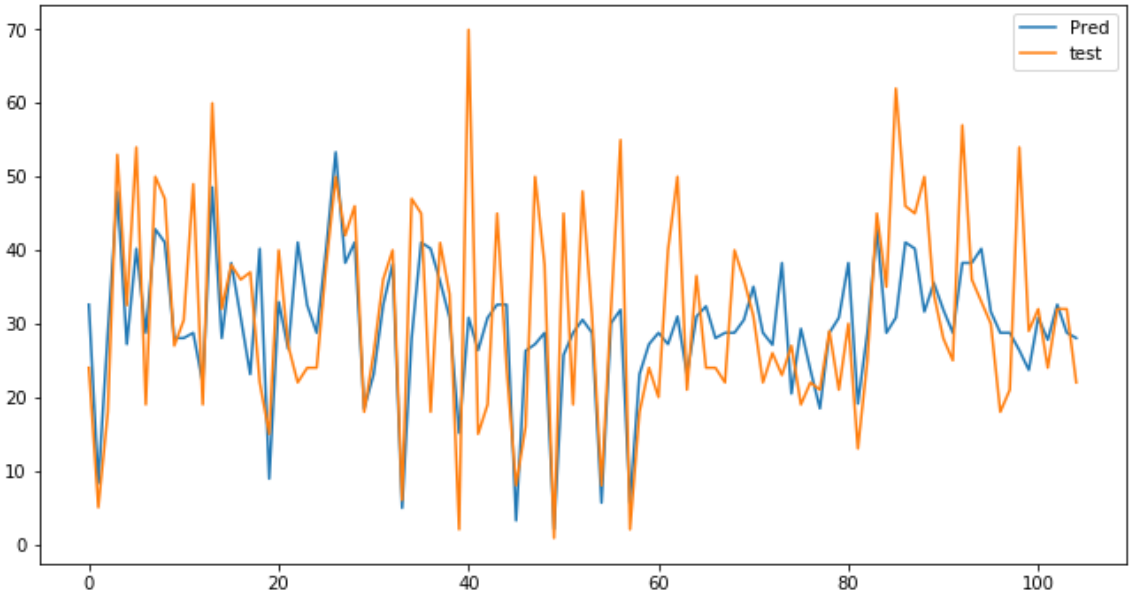
- bins the continuous variable (Fare, Age) in a way so that each feature will be in an increasing/decreasing order
- extracting titles from names, categorize the titles
- combining sibsp and parch into Family feature
- Create (n-1) dummy variable for some features
- Drop some futile features(by experiment)

Missing Value Prediction:

First Try : Without carefully transforming the data, I used polynomial regression to do the fit(The best one after empirical tests), achieving 190 MSE and 10.8 MAE



Second Try: After transforming the data(bins the continuous variable and putting labels in an increasing way, extracting titles from names, combining sibsp and parch into Family feature), all of these approaches made the correlations to ages more clear. Since there is a notable difference for different level within each feature, I decided to use RandomForest to ''Classify" the data and to compute the missing ages. And the result is much better, MSE = 108, MAE = 7.8



Confirming the Correlations:

- **Age:**
  ('baby','child','teenager','young','mid-age','over-50','senior')
  The younger, the better chance to survive

Survived

Age	
1.0	0.658537
2.0	0.404762
3.0	0.445946
4.0	0.322751
5.0	0.412371
6.0	0.385965
7.0	0.125000

- **Family Size:**
  '0' represents family size > 4, others represents the family actual size
  Survival rate increase with family size till 4, too large family size will result much lower survival rate

Survived

Family	
0.0	0.161290
1.0	0.303538
2.0	0.552795
3.0	0.578431
4.0	0.724138

- **Sex:**
  '0' represents Female, '1' represents Male
  Female has far more chance to survive than Male

Survived

Sex	
0.0	0.742038
1.0	0.188908

- **Fare:**
  The bigger the number, the more expensive the ticket is
  Expense of ticket has a strong correlation with survivals

Bins_Fare	Survived
1.0	0.197309
2.0	0.303571
3.0	0.441048
4.0	0.600000

- **Title(Refer):**
  Titles might has overlapped information with Sex and Age, but since we already bins Age, overfitting will not be very likely in this case

refer	Survived
Master	0.575000
Miss	0.701087
Mr	0.156673
Mrs	0.796875
Others	0.318182

- **Pclass:**
  Generally, the smaller the number is, the more expensive the ticket is. Same reason, I have binned the Fare, so overfitting will not occur

Pclass	Survived
1	0.629630
2	0.472826
3	0.242363

Neural Network:

Used 3 hidden layers architecture with units to be (input unit + output unit)/ 2 achieving 84% testing accuracy(Not Optimistically Biased)

Here is the prediction on the missing survival information, this result achieved 81% accuracy rate rated by Kaggle. 0 represent not survived, 1 represent survived.

ID	Survived	922	0	953	0
892	0	923	0	954	0
893	1	924	1	955	1
894	0	925	1	956	1
895	0	926	0	957	0
896	1	927	0	958	0
897	0	928	1	959	0
898	1	929	1	960	0
899	0	930	0	961	1
900	1	931	0	962	1
901	0	932	0	963	0
902	0	933	0	964	1
903	0	934	0	965	0
904	1	935	1	966	1
905	0	936	1	967	0
906	1	937	0	968	0
907	1	938	0	969	1
908	0	939	0	970	0
909	0	940	1	971	1
910	1	941	1	972	1
911	1	942	0	973	1
912	0	943	0	974	0
913	1	944	1	975	0
914	1	945	1	976	0
915	0	946	0	977	0
916	1	947	0	978	1
917	0	948	0	979	0
918	1	949	0	980	0
919	0	950	0	981	1
920	0	951	1	982	1
921	0	952	0	983	0

984	1	1024	0	1064	0
985	0	1025	0	1065	0
986	0	1026	0	1066	0
987	0	1027	0	1067	1
988	1	1028	0	1068	1
989	0	1029	0	1069	0
990	1	1030	1	1070	1
991	0	1031	0	1071	1
992	1	1032	0	1072	0
993	0	1033	1	1073	1
994	0	1034	0	1074	1
995	0	1035	0	1075	0
996	1	1036	0	1076	1
997	0	1037	0	1077	0
998	0	1038	0	1078	1
999	0	1039	0	1079	0
1000	0	1040	0	1080	0
1001	0	1041	0	1081	0
1002	0	1042	1	1082	0
1003	1	1043	0	1083	0
1004	1	1044	0	1084	0
1005	1	1045	1	1085	0
1006	1	1046	0	1086	1
1007	0	1047	0	1087	0
1008	0	1048	1	1088	1
1009	1	1049	1	1089	0
1010	0	1050	0	1090	0
1011	1	1051	1	1091	0
1012	1	1052	1	1092	1
1013	0	1053	1	1093	1
1014	1	1054	1	1094	0
1015	0	1055	0	1095	1
1016	0	1056	0	1096	0
1017	1	1057	1	1097	0
1018	0	1058	0	1098	0
1019	1	1059	0	1099	0
1020	0	1060	1	1100	1
1021	0	1061	0	1101	0
1022	0	1062	0	1102	0
1023	0	1063	0	1103	0

1104	0	1144	0	1184	0
1105	1	1145	0	1185	0
1106	0	1146	0	1186	0
1107	0	1147	0	1187	0
1108	1	1148	0	1188	1
1109	0	1149	0	1189	0
1110	1	1150	1	1190	0
1111	0	1151	0	1191	0
1112	1	1152	0	1192	0
1113	0	1153	0	1193	0
1114	0	1154	1	1194	0
1115	0	1155	1	1195	0
1116	1	1156	0	1196	1
1117	1	1157	0	1197	1
1118	0	1158	0	1198	1
1119	1	1159	0	1199	0
1120	0	1160	0	1200	1
1121	0	1161	0	1201	1
1122	0	1162	0	1202	0
1123	1	1163	0	1203	0
1124	0	1164	1	1204	0
1125	0	1165	1	1205	1
1126	0	1166	0	1206	1
1127	0	1167	1	1207	0
1128	0	1168	0	1208	0
1129	0	1169	0	1209	0
1130	1	1170	0	1210	0
1131	1	1171	0	1211	0
1132	1	1172	1	1212	0
1133	1	1173	1	1213	0
1134	1	1174	1	1214	0
1135	0	1175	1	1215	0
1136	0	1176	1	1216	1
1137	0	1177	0	1217	0
1138	0	1178	0	1218	1
1139	0	1179	0	1219	0
1140	1	1180	0	1220	0
1141	1	1181	0	1221	0
1142	1	1182	0	1222	1
1143	0	1183	1	1223	0

1224	0	1253	1	1282	0
1225	1	1254	1	1283	1
1226	0	1255	0	1284	1
1227	0	1256	1	1285	0
1228	0	1257	0	1286	0
1229	0	1258	0	1287	1
1230	0	1259	1	1288	0
1231	1	1260	1	1289	1
1232	0	1261	0	1290	0
1233	0	1262	0	1291	0
1234	0	1263	1	1292	1
1235	1	1264	0	1293	0
1236	0	1265	0	1294	1
1237	0	1266	1	1295	0
1238	0	1267	1	1296	0
1239	1	1268	1	1297	0
1240	0	1269	0	1298	0
1241	1	1270	0	1299	1
1242	1	1271	0	1300	1
1243	0	1272	0	1301	1
1244	0	1273	0	1302	1
1245	0	1274	0	1303	1
1246	1	1275	1	1304	1
1247	0	1276	0	1305	0
1248	1	1277	1	1306	1
1249	0	1278	0	1307	0
1250	0	1279	0	1308	0
1251	1	1280	0	1309	1
1252	0	1281	0		