Crash Data ML

Brad Burkman

12 March 2021

Contents

1 Reworking the Titanic Notebook

The Titanic notebook is two years old, and I noticed as I read through it that several times the output had warnings that parts of the code were outdated. I sometimes got different results. All of the data prep results were the same. The differences started when I got to the logistic regression coefficients. My results on the right, Kaggle's on the left.

	Feature	Correlation
1	Sex	2.201057
5	Title	0.406027
4	Embarked	0.276628
6	IsAlone	0.185986
7	Age*Class	-0.050260
3	Fare	-0.071665
2	Age	-0.469638
0	Pclass	-1.200309

	Feature	Correlation
1	Sex	2.201527
5	Title	0.398234
2	Age	0.287163
4	Embarked	0.261762
6	IsAlone	0.129140
3	Fare	-0.085150
7	Age*Class	-0.311200
0	Pclass	-0.749007

I saw the same differences in the scores of the different models.

	Model	Score
3	Random Forest	86.64
8	Decision Tree	86.64
1	KNN	84.06
0	Support Vector Machines	82.83
2	Logistic Regression	81.37
7	Linear SVC	79.46
5	Perceptron	79.35
4	Naive Bayes	76.88
6	Stochastic Gradient Descent	76.66

	Model	Score
3	Random Forest	86.76
8	Decision Tree	86.76
1	KNN	84.74
0	Support Vector Machines	83.84
2	Logistic Regression	80.36
7	Linear SVC	79.12
6	Stochastic Gradient Decent	78.56
5	Perceptron	78.00
4	Naive Bayes	72.28

2 Models Scores for Crash Data

Here are the results of the different models for the Crash Data. I turned off the SVM calculation because it was just taking too long for a preliminary snapshot.

	Model	Score
3	Random Forest	99.71
8	Decision Tree	99.71
1	k-Nearest Neighbors	99.59
2	Logistic Regression	99.58
5	Perceptron	99.58
6	Stochastic Gradient Descent	99.58
7	Linear SVC	99.58
4	Naive Bayes	95.82
0	Support Vector Machines	0.00