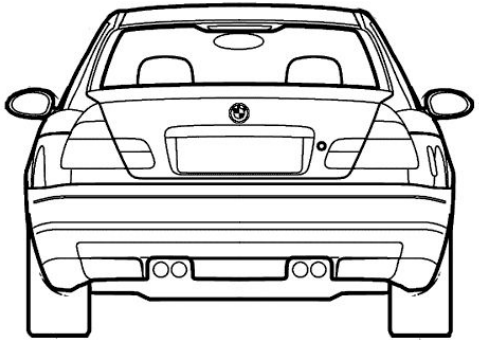
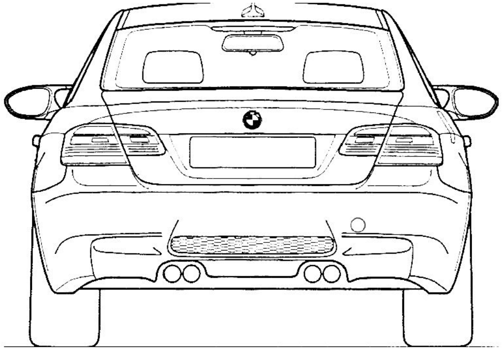
Who ‘e’s that?





By:
Richard Ling
4/24/20

Agenda

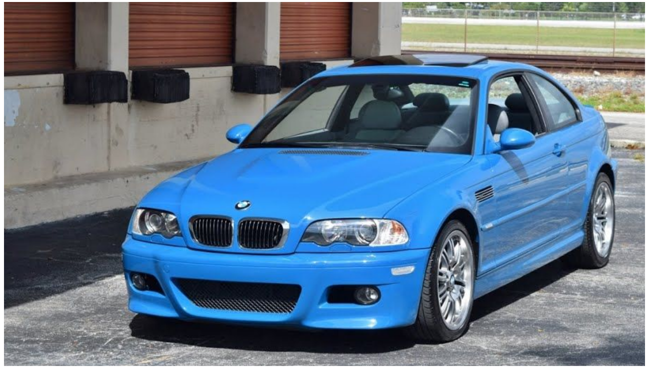
- Overview
- Data Setups
- Models
- Results
- Conclusion

They are e46 and e90s M3 from BMW!

What is 'e'? Code name for 3-Series from BMW

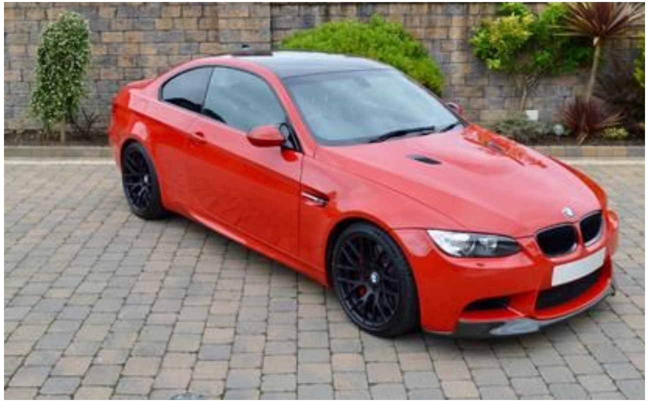
-BMW e46 (1997 - 2006)
Model: M3, 330c(i), 325c(i)

-BMW e90s (2007 - 2013)
Model: M3, 335i, 335is, 328i



Similarities:

Fun driving experience, common radiator problem, fuel pump issues, electronics



Goal: Classify between e46 and e90

Data setup

Gather Data:

- E46: ~2000 posts
- E90: ~1900 posts
- Total: ~ 3900 posts

Cleaning:

- Remove links, punctuations, HTML artifacts
- Result: 2400 post total (1200 for each)

Preprocessing:

- Combine title and selftext as main feature
- Lemmatize the text.

Modeling. Look at me!

Use both CountVectorizer and TfidfVectorizer

1) Logistic Regression
2) K-Nearest Neighbor
3) Navie's Bayes Multinominal
4) Decision Tree

Results

Logistic Regression:

CountVectorizer :

Accuracy = 86%

Total False = 85

TfidftVectorizer:

Accuracy = 85%

Total False = 86

K-Nearest Neighbor:

CountVectorizer :

Accuracy = 58%

Total False = 252

TfidftVectorizer:

Accuracy = 76%

Total False = 143

Navie's Bayes Multinominals:

CountVectorizer :

Accuracy = 84%

Total False = 95

TfidftVectorizer:

Accuracy = 83%

Total False = 100

Decision Tree (Base):

CountVectorizer :

Accuracy = 79%

Total False = 126

TfidftVectorizer:

Accuracy = 83%

Total False = 104

Decision Tree w/ Optimization:

CountVectorizer :

Accuracy = 81%

Total False = 118

TfidftVectorizer:

Accuracy = 81%

Total False = 114

Best Model!

Logistic Regression:

CountVectorizer :

Accuracy = 86%

Total False = 85

TfidftVectorizer:

Accuracy = 85%

Total False = 86

Top coefficients for e46 and e90:

e46 (0)	CVec	TFidf
e46	-1.56	-12.38
330ci	-0.658	-5.70
325ci	-0.485	-4.29
330i	-0.417	NAN
zhp	-0.416	-3.68

e90 (1)	CVec	TFidf
e90	1.31	10.39
335i	0.846	6.527
e92	0.711	5.54
328i	0.650	NAN
2011	0.632	5.437

Results - Misclassifications

Logistic Regression:

-Can’t recognized older model years

-post with very general words.

i.e: cooling leaks, led not working

KNearest Neighbors:

- Can’t recognize the model years and key words

I.e: model year, head gasket

Navie’s Bayes Multinomial:

-Can’t recognized some older model years

Decision Tree:

-With very generic words

i.e. temperature sensor, windshield crack

-Couldn’t get all the keyword and model years.

For all models (total 6 / 600):

-post that were short, very general words and can apply to any model or car.

Example: 'muffler delete opinions yay or nay '

Conclusion

The best model goes to:

- **Logistic Regression w/CountVectorizer and regularization. 86% accuracy.**

Next step/improvement:

- Go back further back for more model years

- Filter out really short post or really short questions.

- Try it on few other models. Such as SVM, Bagging

Last but not least…..

Make your pick and own one!!

e46

e92

f80

Thank you!

Comments / Questions?