Data Mining Project

Notes:

* We need a cost benefit matrix to evaluate models on in the future
* We don’t really care about purchase date since it is irrelevant given that we have the number of years how old the car is, which is much more relevant to the target variable

Variables that don’t matter:

* Purchase Date
  + Irrelevant since we have how old the car is
* RefID, BYRNO
  + Not necessary at all since these are unique identifying values for instance and buyer
* VehYear
  + Irrelevant since we have how old the car is
  + Would only matter if for a certain manufacturer and year there were problems with car, but theoretically should have no impact. Even if this were the case, it would be an anamoly and noise, and we wouldn’t want to fit our data to it.
* *Model*
  + *1063 different distinct values*
  + *Too low level of a measure to really gain any information on*
  + *When using the a Rank+InfoGain Evaluator, Model comes up very low with the target of isBadBuy*
* *Trim, SubModel*
  + *All similar to make in terms of too may distinct values and the InfoGain Evaluator ranked them very low*
* IsOnlineBuy
  + No information gain
* Transmission
  + No to little Information Gain
* Auction, VNST, Color, Nationality , TopThreeAmerican
  + No to little Information Gain