Collinearity - What it means, Why its bad, and How does it affect other models?

Elliott Saslow Follow
Jul 11, 2018

Questions:

> What is a collinearity or multicollinearity? Why is it bad? What does it look like?
>
> How does it affect our results?
>
> Does it affect decision trees?

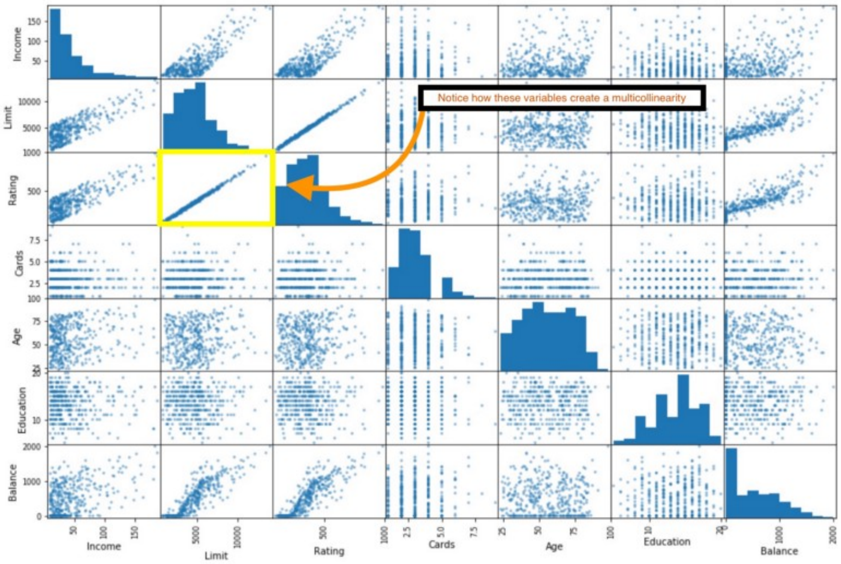
1 In statistics, **multicollinearity** (also **collinearity**) is a phenomenon in which one feature variable in a regression model is highly linearly correlated with another feature variable.

A **collinearity** is a special case when two or more variables are exactly correlated.

This means the regression coefficients are not uniquely determined. In turn it hurts the interpretability of the model as then the regression coefficients are not unique and have influences from other features. **The ability to interpret models is a key part of being a Data Scientist.**

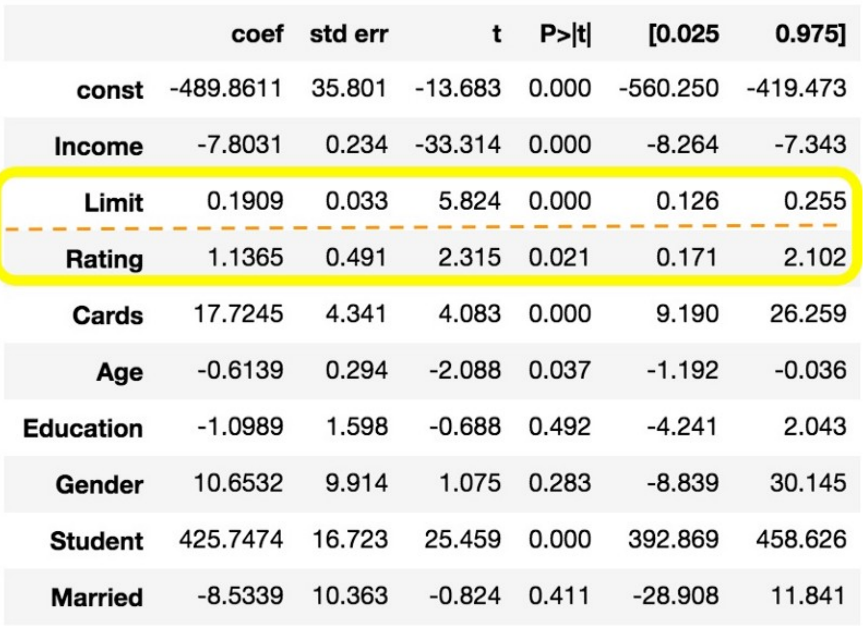
Regardless, if you are just in the business of predicting, you don’t really care if there is a collinearity, but to have a more interpretable model, you should avoid features that have a very high (~R^2 > .8) being contained in the features.

Below is an image of the data set I am working with, the shows scatter plots of many of the variables in the dataset. Notice how **Limit** and **Rating** are so clearly highly correlated. This implies a multicollinearity and takes away from our ability to interpret the beta coefficients from both.



Scatter matrix of variables

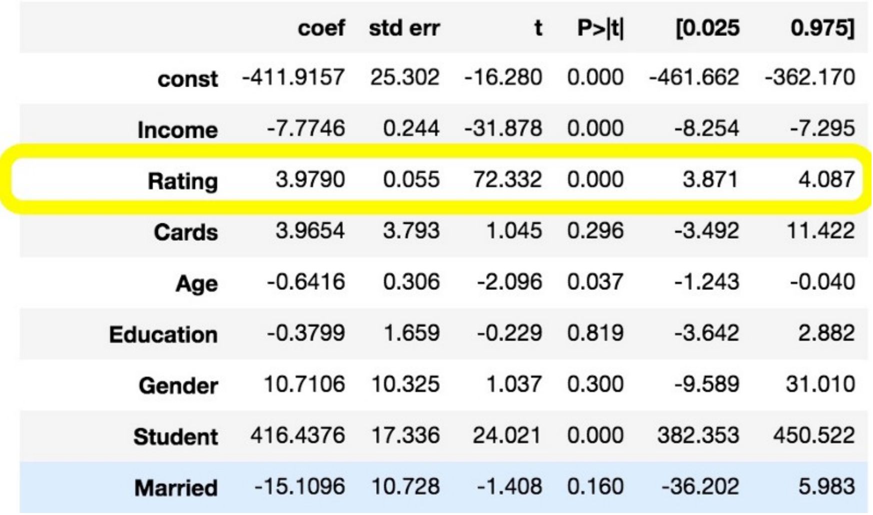
So now, if we use linear regression to predict the balance of each person, we can look at our beta coefficients. Unfortunately because of the multicollinearity it becomes harder to understand what is going on:



	coef	std err	t	P>|t|	[0.025	0.975]
const	-489.8611	35.801	-13.683	0.000	-560.250	-419.473
Income	-7.8031	0.234	-33.314	0.000	-8.264	-7.343
Limit	0.1909	0.033	5.824	0.000	0.126	0.255
Rating	1.1365	0.491	2.315	0.021	0.171	2.102
Cards	17.7245	4.341	4.083	0.000	9.190	26.259
Age	-0.6139	0.294	-2.088	0.037	-1.192	-0.036
Education	-1.0989	1.598	-0.688	0.492	-4.241	2.043
Gender	10.6532	9.914	1.075	0.283	-8.839	30.145
Student	425.7474	16.723	25.459	0.000	392.869	458.626
Married	-8.5339	10.363	-0.824	0.411	-28.908	11.841

is **Limit** or **Rating** driving the results?

Both limit and rating have positive coefficients, but it is hard to understand if the balance is higher because of the rating or is it because of the limit? I think the driving influencer here is rating, because with a high rating, you achieve a higher credit. So I would remove **Limit** to get a true idea of how the rating affects the balance.



	coef	std err	t	P>|t|	[0.025	0.975]
const	-411.9157	25.302	-16.280	0.000	-461.662	-362.170
Income	-7.7746	0.244	-31.878	0.000	-8.254	-7.295
Rating	3.9790	0.055	72.332	0.000	3.871	4.087
Cards	3.9654	3.793	1.045	0.296	-3.492	11.422
Age	-0.6416	0.306	-2.096	0.037	-1.243	-0.040
Education	-0.3799	1.659	-0.229	0.819	-3.642	2.882
Gender	10.7106	10.325	1.037	0.300	-9.589	31.010
Student	416.4376	17.336	24.021	0.000	382.353	450.522
Married	-15.1096	10.728	-1.408	0.160	-36.202	5.983

Notice Rating is higher

Here you can now see that **Rating** has a higher impact than **Limit + Rating** did before. This is more interpretable to those who do not understand the math.

2 Even so, between the two models, the model with both variables (Limit & Rating) performed better (by R² scoring). This leads to a discussion on why we care in the first place. We want to use these models to help us to understand the world around us, and figure out where to take our data exploration. **Therefore when applying linear regression, you may want to use different models for prediction and one for interpretation/inference.**

This same concept can be applied with a **Collinearity** such as getting the dummy variables for **Ethnicity.** In this case by keeping all of the dummy variables, you lose the ability to interpret how each variable affects the results. With a Collinearity, removing a column **does not affect results.**

3 Finally, since these issues affect the interpretability of the models, or the ability to make **inferences** based on the results, we can safely say that a multicollinearity or collinearity will not affect the results of predictions from decision trees. During inference from the decision tree models though, it is important to take how each feature may be affected by another into account to help make valuable business decisions.