**Course 5 Task 3**

**Michael DeGraff**

**Customer Default Identification Report that addresses:**

**Problem:**

An increase in customer default rates is bad for Credit One since its business is approving customers for loans in the first place. This is likely to result in the loss of Credit One's business customers.

Questions to Investigate:

* How do you ensure that customers can/will pay their loans?
* Can we approve customers with high certainty?

As you have already learned, we are limited in terms of what we can control and what the data will be able to tell us. Specifically, we need to remember:

* We cannot control customer spending habits
* We cannot always go from what we find in our analysis to the underlying "why"

We these limitations, we must focus on the problems we can solve:

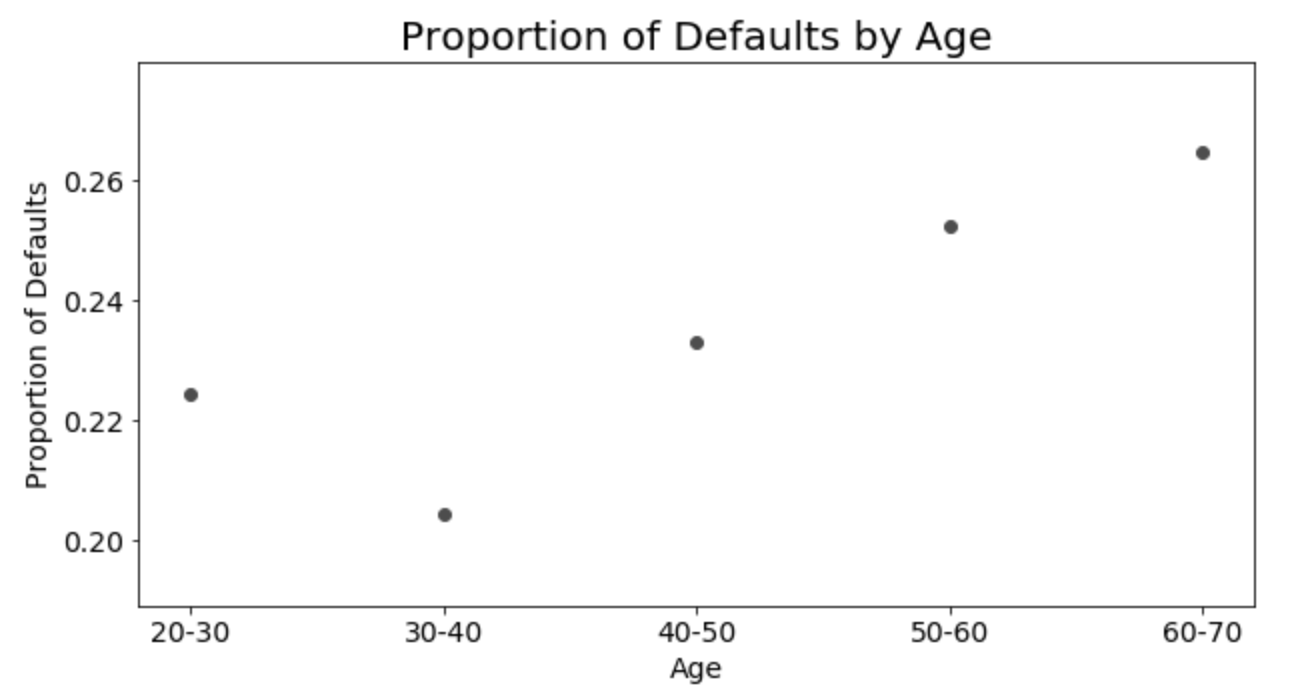
* 1. Which attributes in the data can we deem to be statistically significant to the problem at hand?
  2. What concrete information can we derive from the data we have?
  3. What proven methods can we use to uncover more information and why?

Which attributes in the data can we deem to be statistically significant to the problem at hand?

Digging into the data reveals that the given information does not provide a great deal of insight into the likelihood of a customer default. Though there are a couple of insights.

*Age*

Figure 1

As you might suspect, as age increases the likelihood of default decreases. However, there is a dip from the 20-30 age group to the 30-40 folks (see figure 1).

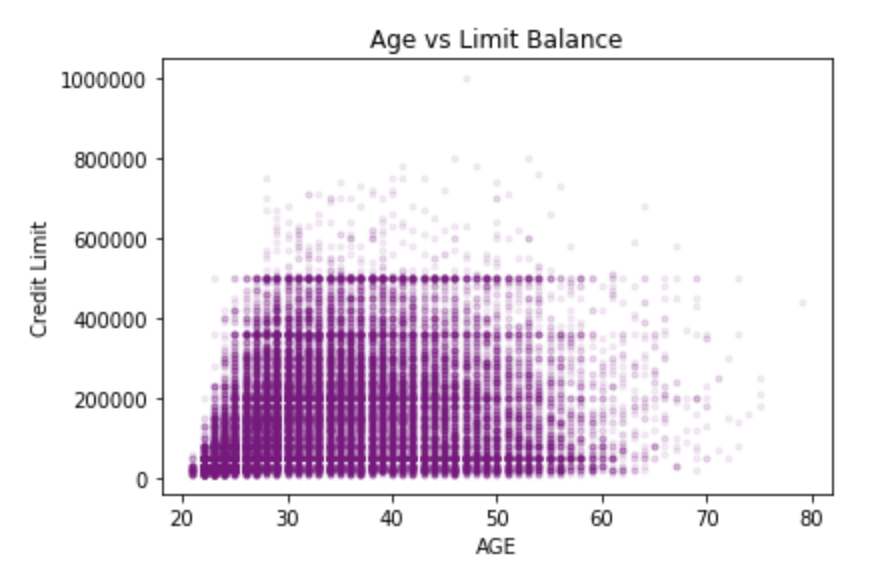
However, once you consider the relatively low credit limits for 20 year old clients, it makes more sense that fewer young adults would default. Figure 2 shows the credit limit vs age from the data set used.

Figure 2

*Credit Limit*

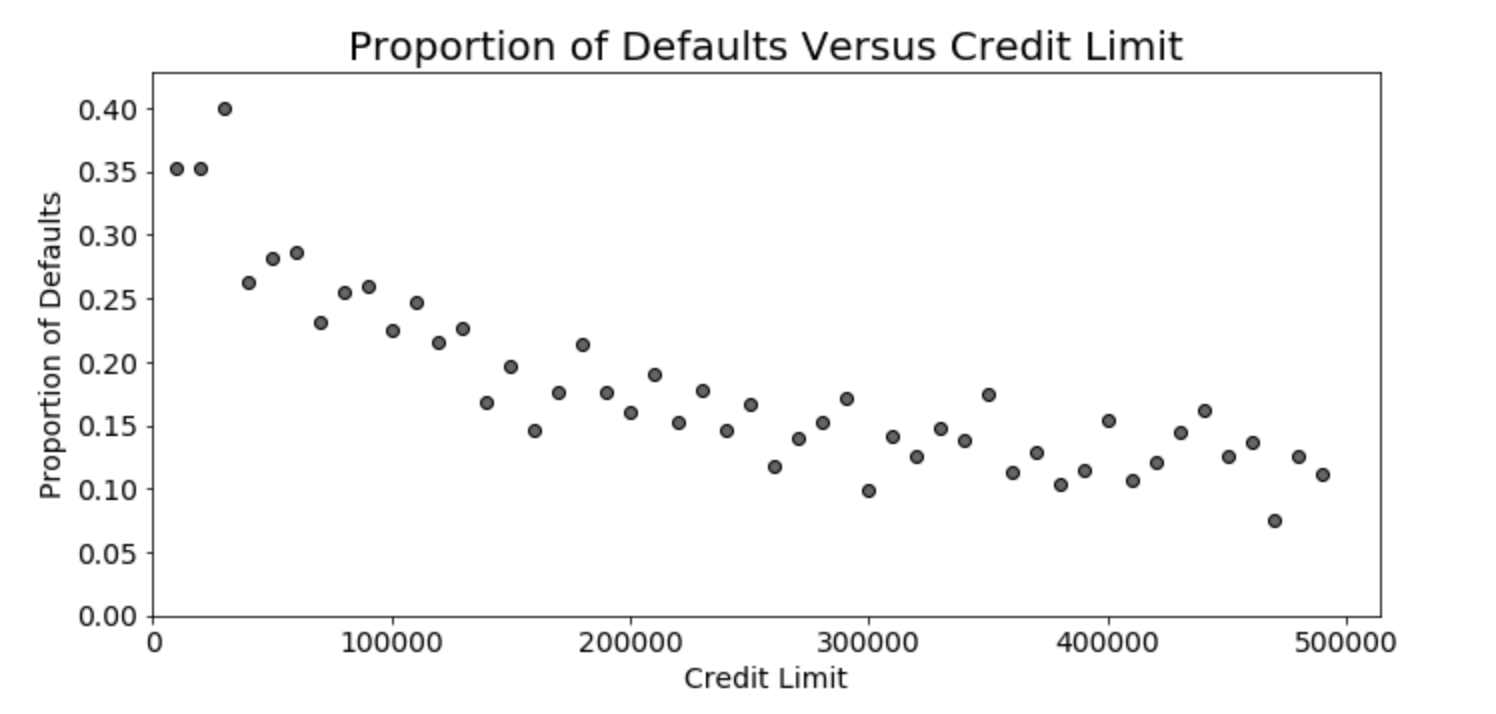


Figure 3

Speaking of credit limits, this also plays a role in the likelihood of default. As age increases, the proportion of defaults steadily declines.

Figure 4

*Modeling*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **SVM** | **RF** | **KNN** |
| Accuracy | 0.78 | 0.83 | 0.78 |
| Kappa | 0.36 | 0.32 | 0.40 |
| Precision | 0.81 | 1.00 | 0.54 |

Given the initial insights related to credit limit and age, we looked more closely at the defaults by creating a few models. We ran several estimators, but we narrowed it down to three: Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbor (KNN). As figure 4 shows, the accuracy scores were similar across the board, as well as Kappa. The differentiator ended up being the Precision score, which tells us how often predicted defaults are correct. While RF appears to be best, it is clear that we have a case of overfit. SVM precision is significantly better than KNN. However, in all three cases, the models did not do better than chance when compared to the actual results (ground truth). Our ultimate conclusion is to find additional attributes that might provide more insight into future default status.