**PROJECT**

Explore and Summarize Data

A part of the Data Analyst Nanodegree Program

* [**PROJECT REVIEW**](https://review.udacity.com/)
* [**CODE REVIEW**](https://review.udacity.com/)
* [**NOTES**](https://review.udacity.com/)

**SHARE YOUR ACCOMPLISHMENT**

Requires Changes

**3 SPECIFICATIONS REQUIRE CHANGES**

Kindly comment on the predictive model used as this topic is new to me and was missed during previous review. Thanks!

Apology on behalf of the previous reviewer for the missing answer.

Your method here is acceptable, looking at all the performance metrics we can see that the model was getting better as each additional variable was added.

I noticed two problems in this model creation process:

Firstly, the use of "accuracy" metric poses a problem since the values of "LoanStatus" were not evenly distributed:

Completed Defaulted Chargedoff

34748 4572 10659

Imagine a dummy classifier that predicts all instances as "Completed", its accuracy would be 34748 / (34748+4572+10659) = 69.5%, better than any of our trained classifiers! A better set of metrics to use here are [precision and recall](http://rushdishams.blogspot.com/2011/03/precision-and-recall.html) which gives you a better measurement of quality of your classifier.

The second problem was the train/test split process. Imagine an extreme case where all of the observations with "Chargedoff" LoanStatus are at the end of the dataset. In that case, almost all of the test dataset would have "Chargedof"f LoanStatus. To anticipate this scenario, you may decide to randomize the observations before splitting the dataset, but the risk is still there, especially with the uneven distribution of LoanStatus values i.e. we expect that most of the test and training dataset to have "Completed" LoanStatus.

To handle this issue, we may do a [k-fold cross-validation](https://www.cs.cmu.edu/~schneide/tut5/node42.html) where all observations will be used evenly. It splits the dataset into k folds, then iterates the training and testing k-times. In each iteration, one subset of the dataset used as test data and the rest for training. At the end of the validation process, the performance score is then averaged. This method allows all of the LoanStatus labels to be used evenly, which in the end would produce a model that better predicts new features it has never seen before.

Now, I do not generally use R for other than data analysis. Personally, I find Python's module scikit-learn to have a more consistent pipeline for machine learning. Hence, I am afraid I am not able to advise you on which packages to use in R to mitigate the problems above. On the other hand, there is no need to worry since you will learn all about this in the next course "Intro to Machine Learning". I think you will find this course really interesting.

There are only a couple of minor issues in the report. Good luck with your next submission!

**Code Functionality**

**All code is functional (e.g. No Error is produced and RMD document is not prevented from being knit.)**

**The project almost never uses repetitive code where a function would be more appropriate. The code references variables by name instead of using constants or column numbers.**

**Project Readability**

**All complex code is adequately explained with comments. It is always clear what the code is doing and how and why any unusual coding decisions were made.**

**The code uses formatting techniques in a consistent and effective manner to improve code readability. All lines are shorter than 80 characters.**

**Markdown syntax is used in the RMD file to improve readability of the knitted file.**

Good job adding line breaks as previously suggested.

(Optional) Since the report is quite long, I think a table of content would be a useful feature to incorporate here. The following Markdown syntax can be added on top of the document to create a dynamic table of contents that can also act as a navigator:

---

title: "Analysis Title"

author: Your Name

date: [Month] [date], [year]

output:

html\_document:

toc: true

toc\_depth: 3

toc\_float: true

---

**Quality of Analysis**

**The project appropriately uses univariate, bivariate, and multivariate plots to explore most of the expected relationships in the data set.**

Good job handling outliers as previously suggested.

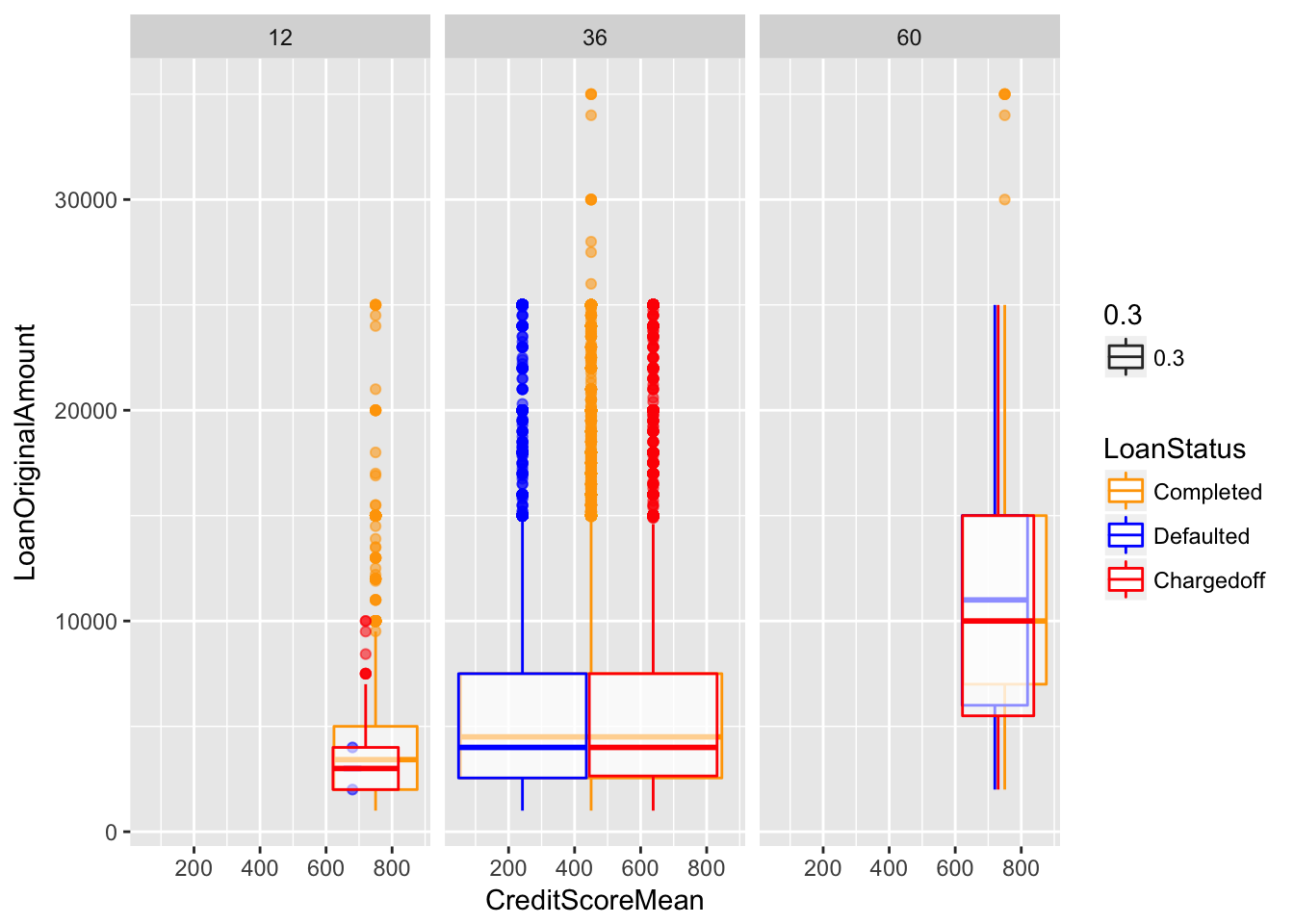
**Questions and findings are placed between blocks of R code regularly so it is clear what the student was thinking throughout the analysis.**

**Reasoning is provided for the plots made throughout the analysis. Plots made follow a logical flow. Comments following plots accurately reflect the plots’ contents.**

**The project contains at least 20 visualizations. The visualizations are varied and show multiple comparisons and trends. Relevant statistics (e.g. mean, median, confidence intervals, correlations) are computed throughout the analysis when an inference is made about the data.**

**Visualizations made in the project depict the data in an appropriate manner that allows plots to be readily interpreted. Choice of plot type, variables, and aesthetic parameters (e.g. bin width, color, axis breaks) is appropriate.**

Good job improving the plots as previously suggested. There is this one plot that I think can be further improved:



First of all, the legend "0.3" is not needed here as it does not add new information. It can be removed by placing alpha = 0.3 outside of aes parameter.

Secondly, as the plot shows two continuous variables, one for each axis, a scatterplot would be a better plot type to use here, probably with geom\_smooth geom to help readers identifying the trend for each loan status.

**Final Plots and Summary**

**The project includes a Final Plots and Summary section containing three plots and commentary. All plots in this section reflect what has been explored in the main body of the analysis.**

**The plots are well chosen and the plots fulfill at least 2 of the criteria. The plots are varied and reveal interesting trends and relationships.**

**All plots have appropriately selected variables and are plotted in a way that accurately conveys the data/information (i.e findings in Final Plot 1 do not depend on the findings of Final Plot 2).**

Good job improving final plots as suggested. Notice that final plot three has the redundant legend "0.3" there.

**All plots are labeled appropriately (axis labels, plot titles, axis units) and can be read and interpreted easily. Plots are scaled appropriately.**

**Missing titles**

Plot titles need to be added to all of the final plots. This is important so when the plot image is saved the title remains there and the plot does not lose its context.

**Add units of measurements**

Currency name needs to be added for "LoanOriginalAmount" variable, and readers need to know what "12" , "36", and "60" are in final plot three.

I understand it does sound picky, but adding a unit of measurement even though it is obvious is *super important*. [NASA can tell you more about it](http://articles.latimes.com/1999/oct/01/news/mn-17288):

NASA lost its $125-million Mars Climate Orbiter because spacecraft engineers failed to convert from English to metric measurements when exchanging vital data before the craft was launched, space agency officials said Thursday.

A navigation team at the Jet Propulsion Laboratory used the metric system of millimeters and meters in its calculations, while Lockheed Martin Astronautics in Denver, which designed and built the spacecraft, provided crucial acceleration data in the English system of inches, feet and pounds.

As a result, JPL engineers mistook acceleration readings measured in English units of pound-seconds for a metric measure of force called newton-seconds.

In a sense, the spacecraft was lost in translation.

**The reasoning and findings from each plot are explained and the text about each plot is descriptive enough to stand alone. Comments reflect the contents of the plots that they are associated with.**

**Reflection**

**The project includes a Reflection section discussing the analysis performed.**

**The section reflects on how the analysis was conducted and reports on the struggles and successes throughout the analysis. The section provides at least one idea or question for future work. The section explains any important decisions in the analysis and how those decisions affected the analysis.**