Lending Club
Capstone Project 1

Problem Statement:

Will a borrower default on their loan and have their loan charged off?

How the data was obtained, cleaned, and wrangled:

OBTAIN DATA:

Data is from LendingClub: https://www.lendingclub.com/info/download-data.action

EXPLORE DATA:

For the data wrangling I started with a smaller subset of the entire data to begin exploring and understanding general aspects of the data.

After pulling the smaller subset of data from a CSV file, I compared the first 5 rows of data to the data types of each column. Two issues jumped out from that quick analysis. The first was date columns were imported as objects instead of dates. The second was a few columns that were labeled as object which could be converted to floats.

FORMAT DATA TO NUMBERS WHEN POSSIBLE:

To convert the dates I re-imported the data and used a parameter to ensure that the dates were read in appropriately.

There were two columns that were percents but when imported they were converted to strings because of the percent sign. I simply stripped off the percent signs and converted them to floats.

The loan term was entered as either “36 months” or “60 months”. Both of those columns could be improved by removing the word “months” and changing them to floats. After converting the data I changed the column name to better reflect that the data was in months.

Similarly, the employee length column was labeled in years which could be improved if turned into a float. However, there were two unique values that needed some interpretation. Some rows said, “10+ years”. I altered those to “10 years” as there was no other label for 10 years and I felt it would continue to give a good representation of the data. There were other rows that said “< 1 year”. For those I felt that 0.5 years was a good approximation. After that the year label was stripped off, the data was converted to a float, and the column label was changed to reflect that the data was in years.

IMPORT FULL DATASET:
At this point I was feeling more comfortable with all the data. I read in the entire set of available data which was over 2 million rows and 144 columns. I made the changes discussed to get the data types in good order.

DELETE EMPTY ROWS:
There were some rows that were missing loan amounts. After further investigation it turned out there was no data in those rows. They were holding information for the spreadsheet and do not apply to the current analysis. The rows were deleted.

CLEAN LOAN STATUS:
The column Loan Status had some text added in places that described if the loan met the credit policy. This is essentially adding another set of data into the Loan Status column. I created a new column to hold the additional data then removed it from the Loan Status column.

In general we need to have loan status on every loan. Outside of that, I do not think we need to all rows to be completely filled in to try and predict loans that have become delinquent.

DELETE EMPTY COLUMNS:
With that in mind, there were three columns that had no information at all. Those were removed.

HANDLE OUTLIERS:
Looking through the data with .describe() I noticed that the maximum annual income seemed very high. I looked at the top few annual salaries and there were two that were significantly larger than the others. One was listed as $110 million dollars a year and the next one at $61 million dollars a year. The third highest was a “reasonable” $11 million dollars a year. A scatter plot made the disparity easy to see. I made what seemed to be reasonable corrections to the two outlier salaries.

There were still 676 borrowers with annual salaries over $1 million a year out of the 2,250,000 total borrowers. This seemed unlikely but not impossible, so I decided to keep the data.

IMPOSSIBLE DTI RATIO:
Finally, two borrowers had a negative debt to income ratio. This is not possible. Those values were changed to Null as it is difficult to ascertain why the values were entered as negative.

The data feels fairly clean at this point and it would be a good time to transition from wrangling of the data to more visualizations of the data.

Github link to Jupyter Notebook:
https://github.com/mmiller124/Springboard/blob/master/LendingClub/Capstone%20Project%201%20Data%20Wrangling.ipynb

Exploratory Analysis:

FURTHER REDUCE DATA:
Pull out the relevant data that only includes the Fully Paid, Charged Off, and Default loans.

WHO, WHY, AND HOW MUCH:
Look at what occupations take out the most loans, why they are taken out, and how big the loans are. Most people fell under the label of teacher, nurse, or manager. Loans were taken out overwhelmingly for debt consolidation. The loans range in size from 500 dollars up to 40,000 dollars with most being around 10,000 dollars.

RELEVANT OBSERVATIONS:
Going through the data one column at a time, some columns seemed to be more relevant. I narrowed my focus from many to just a few columns and viewed the violin plots to see the data. Here are the areas that I wanted to focus:

- Loan Status
- Loan Amount
- Annual Income
- Installment (monthly payment)
- DTI (debt-to-income ratio)
- Credt inquiries in the last 6 monts for the secondary applicant
- Ratio of total currnt balance to high credit/credit limt for all revolving accounts
- Number of charge-offs within last 12 monts at time of application for the secondary applicant

. A heat map gave correlation between these columns but, surprisingly, there was not much.

ANNUAL INCOME:
For statistical analysis I narrowed my focus even more on Annual Income. Histograms and Empirical Cumulative Distribution Functions helped visualize the data with increased focus. The null hypothesis that there is no difference between the Annual Incomes of the Charged Off/Default population and the Annual Incomes of the Fully Paid income resulted in an incredibly small p-value. With more analysis we were able to give a confidence interval for the difference between the average annual income in the two groups. Analysis showed 95% confidence that the difference between the average income of the two groups was between $6912 and $7476.

Machine Learning:

This is a typical Classification model for Data Science as our information can be broken down into two groups, charged off (default) or fully paid. Having some inference to understand what inputs affect the prediction would be helpful, but the bottom line is understanding if borrowers will end up in default or not.

In previous sections the data was cleaned, explored, and statistical analysis was run. Now it is time to make predictions with Machine Learning.

In order to make predictions and use machine learning algorithms, feature encoding is required. Machine learning is based on Linear Algebra so all values must be numeric. Categorical, or non-numeric, features need to be converted to numeric categories. One of the best ways to do this is using Pandas get_dummies.

The variability and interdependence of a dataset can skew the results of a machine learning prediction. We can use a heatmap to give a visual representation of the interdependence of variables in a dataset. Then we can easily drop the variables that are related.

At that point the data can be split into test and training datasets. The training dataset can be further broken down into smaller datasets to provide validation of hyperparamerters.

From there we can start to use some supervised machine learning tools. Using Ridge and Lasso linear regression would be a good starting point to evaluating the data as well as basic Logistic Regression. K-Nearest Neighbors, SVM, CatBoost, and Random Forest also provide strong predictions based on regression. These tools will be trained with the training data and tested with the testing data which it has never seen before.

From here we will have a number of results so we need a way to compare the quality of the predictions. The AUCROC will give us a very good way to compare results from different machine learning tools. Finally, PCA can give us a visual representation of the data in 2 or 3 dimensions that lets us have a feel for how well the prediction worked.

After all these steps are complete we will have a strong prediction algorithm that a business can use to determine if borrowers will default on their loans. This can save money for a business with strong statistics based predictions and can also give them a feel for how often they will get the predictions incorrect before using it in a production environment.