Data Analysis Report

Loan Predictor

by Zhaklin Yanakieva

Table of content

Versioning table

Time period(weeks)	Comments on changes	Version
5 — 7	Initial changes to the document — creating the structure	1.0.0
8 — 10	Filling up with the needed information, such as explaining the graphs, and summarizing	2.0.0
11 — 12	Adding the Data storage solution	3.0.0

Overview:

The data analysis report is to provide the findings of the research that was conducted during the collection of the data for this project.

Background:

In the DataCollectionLedger file is explained more about how and from where the data was collected. In the proposal of the project is explained the goal and for what models the data will be used.

Data storage solution:

An essential part of Machine Learning is the data storage solution for the selected data and machine learning model. In order to accomplish the most efficient manner of working with data during this project, the following tools were used:

- Git Version Control
- Data Version Control (DVC)

Git Version Control:

Git has been a popular tool among programmers and it is so for a reason. It allows tracking changes in any set of files, usually used for coordinating work among programmers collaboratively developing source code during software development.[1]

[1] https://en.wikipedia.org/wiki/Git

Data Version Control (DVC):

Data Version Control is a new type of data versioning, workflow, and experiment management software that builds upon Git (although it can work stand-alone).[2] Using Git and DVC, machine learning teams can version experiments, manage large datasets, and make projects reproducible. By utilizing DVC data will be tracked and stored in an effective and efficient way because the data is accessible from everywhere via internet connection for every contributor.

Summary:

- DVC will create reference files to data versions
- Git will store the DVC files

At the current stage, the dataset has been loaded after integrating it. Changes were made to like cleaning and processing it so as to make it more suitable to work with and acceptable to store.

Purpose of the extraction effort:

Understanding, preparing and cleaning the data require first to know what type of it will be used and for what reason. The purpose of the integration of two csv files in the Loan Predictor is to combine more data that will help for the gathering of different information. In this case, the results will not be only from one source, which states that the predictions will be more accurate for real life.

Data integration:

In the project, I decided to combine data from two different csv files in order to gather more information. After that, data integration was applied in order to gather the data from two different csv files. The method that was used was ‘Union’, which makes me use exactly the same data

[2] https://dvc.org/doc/user-guide/what-is-dvc

characteristics for every column (type, domain and cast) so as to combine the datasets. For the 'Loan Predictor', I decided to follow this technique to create bigger data from which I can conclude properly the results after the training of the models. In this way, the bigger the data, the more to explore, which means I got the opportunity to go through more cases and deduct the final result.

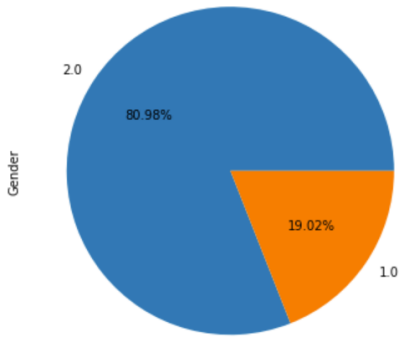
Methods:

There are many ways to retrieve data from a source you selected. Common ways of extracting are from a file (csv or txt), from a database, JSON, API or web-scraping. In my project, I chose to extract the data from two different csv files. I first check each file for the needed information about the loans and then, save the data together.

Results:

```
[ ] df['Gender'].value_counts().plot(kind='pie', autopct='%1.2f%%', figsize=(6, 6))
```

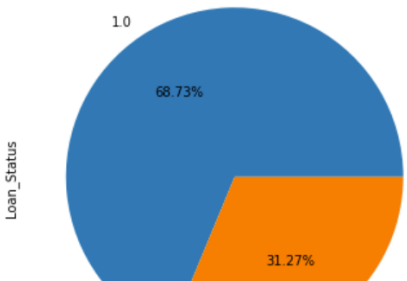
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f99e0c59f50>
```



The pie chart above represents the percentage of the gender value. It is noticeable that the percentage of males who applied for a loan is greater than the one of females.

```
df['Loan_Status'].value_counts().plot(kind='pie', autopct='%1.2f%%', figsize=(6, 6))
```

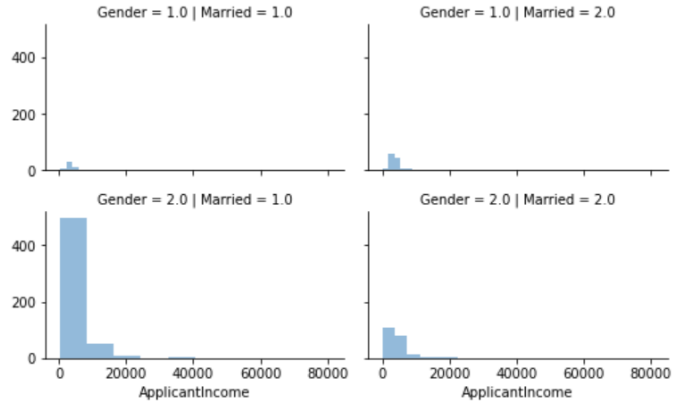
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f99cd11d410>
```



The pie chart above represents the percentage of the loan status. Results show that there are more approved loans that disapproved.

```
grid=sns.FacetGrid(df, row='Gender', col='Married', size=2.2, aspect=1.6)
grid.map(plt.hist, 'ApplicantIncome', alpha=.5, bins=10)
grid.add_legend()
```

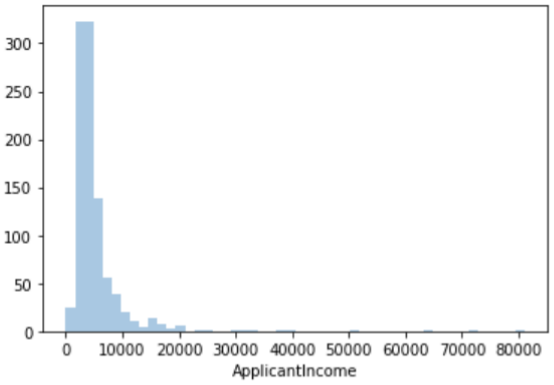
```
/usr/local/lib/python3.7/dist-packages/seaborn/axisgrid.py:316: UserWarning
  warnings.warn(msg, UserWarning)
<seaborn.axisgrid.FacetGrid at 0x7f99cb69e8d0>
```



These histograms display the gender and marriage in accordance to the applicant income. It can be noticed that males have the highest income according to the data. Males that are married have greater income than unmarried male. And the same goes for females.

```
sns.distplot(df.ApplicantIncome,kde=False)
```

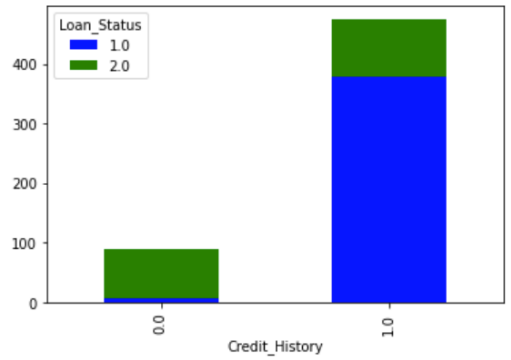
```
/usr/local/lib/python3.7/dist-packages/seaborn/distr
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f99c26
```



The histogram represents that people with better education should normally have a higher income.

```
temp3 = pd.crosstab(df['Credit_History'], df['Loan_Status'])
temp3.plot(kind='bar', stacked=True, color=['blue','green'], grid=False)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f99c23e5490>
```



This diagram outlines that the chances of getting a loan are higher if the applicant has a valid credit history.

Summary:

From the data analysing, it can be concluded that the amount of male applicants is greater than the female ones and they tend to live in the semi suburban areas. Also, there are more positive than negative loan statuses - more approvals. The distributions show that the graduates have more outliers which means that the people with huge income are most likely to be educated. Lastly, males have the highest income according to the data and there are more married males with greater income than unmarried ones. And the same goes for females. Therefore, there is a greater chance for educated and married people to receive a loan than applicants who are not.

From the Modelling analysing, it is concluded that the more accurate model is Random forest than Decision tree. From the evaluation of the three models, the Logistic Regression performed better than the others.

Bibliography:

DVC. (n.d.). https://dvc.org/doc/user-guide/what-is-dvc

Git. (n.d.). https://en.wikipedia.org/wiki/Git