**Paper Title (Something Interesting)**

**Find a relevant dataset from Kaggle- On classification**

Name, [student@bellarmine.edu](mailto:student@bellarmine.edu)

**ABSTRACT**

Up to 150 word summary of your project.

1. **INTRODUCTION**

My dataset has to do with credit card approval. It takes in several variables, and based on those, tells you if the person has been approved for a credit card or not. I decided to use Logistic Regression using Python for my method of classification.

1. **BACKGROUND**
   1. *Data Set Description*

I found the credit card approval dataset on Kaggle when looking for datasets to use. I chose it because it seemed interesting to see how different variables can determine something like approval for a credit card. I think this is very similar to what I will be doing in my career field in that I am going to be taking and processing variables and determining things about one’s insurance.

* 1. *Machine Learning Model*

Logistic Regression in Python is a step-by-step process. First you must import your dataset. Then, you clean up the dataset so that it is ready for analysis. Then you make sure all the categorical variables are numerical. Then, you split the variables into dependent and independent variables (there will only be one dependent variable). Then, you split the dataset into training set and testing set, usually selecting 10 percent or 15 percent or the testing set size. Then, you train the Logistic Regression model on the Testing set using sklearn and the LogisticRegression() function. Then, you are finished and can accurately predict the result.

1. **EXPLORATORY ANALYSIS**

The dataset contains 690 entries with 15, some of which were removed for simplicity of the data.

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| Gender | Integer |
| Age | Float |
| Debt | Float |
| Married | Integer |
| Prior Default | Integer |
| Employed | Integer |
| Credit Score | Integer |
| Income | Integer |
| Approved | Integer |

1. **METHODS**

I prepared the data by scanning it and finding out what I needed to do first. Since there were 15 columns, 14 independent variables, I decided to remove some of them to have less independent variables. I removed 6 of the columns that I deemed least important to the result.

* 1. *Data Preparation*

I prepared the data by scanning it and finding out what I needed to do first. Since there were 15 columns, 14 independent variables, I decided to remove some of them to have less independent variables. I removed 6 of the columns that I deemed least important to the result.

* 1. *Experimental Design*

In my first experiment, I did 15 percent for the testing set size. For my second experiment, I did 10 percent for the testing set size.

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | 80/10/10 split for train, validate, and test |
| 2 | 70/15/15 split for train, validate, and test |

* 1. *Tools Used*

Describe all of the software tools you used to perform your data preparation and model implementation. For example:

The following tools were used for this analysis: Python v3.5.2 running the Anaconda 4.3.22 environment. In addition to base Python, the following libraries were also used: Pandas 0.18.1, NumPy 1.11.3, SKLearn 0.18.1. I needed to pandas to load the dataset onto Python. I needed NumPy for the different arrays used throughout. I used SKLearn in order to obtain the Logistic Regression function for the classifier.

1. **RESULTS**

I got 82 percent accuracy. I got a value of 48 on the true positive, 13 on the false positive, 6 on the false negative, and 37 on true negative.

1. **CONCLUSION**

The results weren’t bad, but they could use a little work. The model was ok, but I was hoping for a little bit higher on the confusion matrix. I think it could be improved by having a better dataset. There may have been some values that made it less accurate.