**Program 2 - Task 1:**

**Data Science with Python 2020**

Guido Rossum, Senior Data Scientist, at Credit One asked us to begin a new data science project using Python. Over the past year Credit One has encountered an increase in the number of customers who have defaulted on loans they secured from various partners, and Credit One, as their credit scoring service, could risk losing business if the problem is not solved right away. The bottom line is Credit One needs a much better way to understand how much credit to allow someone to use.

Credit One enlisted the help of our data science team to design and implement a creative empirically sound solution. It is very important that we understand from the start that this is not a typical data analytics problem as we have been given full authority to solve this problem with whatever tools and methods we need. With that in mind we elected to use Python and a few different libraries to do the heavy lifting for us. We'll be using a few main libraries for the bulk of our work, but will not limit ourselves to only using those - we will investigate other libraries that may contribute to the best solution.

We will define this problem within a framework before using Credit One’s historical data in our own working environment and Python to attempt to solve the problem.

In defining the problem within a data science framework we can understand the differences between what Credit One has previously done with data analytics and what we will do in this project with data science. To complete our work we will use our local programming environment to have access to the tools we need, before finally starting the analysis and solving this problem. A report on our experience and understanding of the problem will follow.

**Outline of the report for our client's management that we will use to create a PowerPoint presentation.**

1. Purpose & Goals -
   1. Credit One has encountered an increase in the number of customers who have defaulted on loans and risks losing business if the problem persists. Credit One requires a better way to understand how much credit to allow someone to use.
   2. We will design and implement a creative empirically sound solution:
      1. To decrease the number of customers who default on loans secured
      2. Map a better understanding of how much credit to allow someone to use.
      3. To identify and predict customer default using credit card historical data provided by Credit One.

As a result Credit One we will be able to predict with reliable accuracy the appropriate amount of credit to extend customers to minimize risk of loss of revenue, an increase in number of loan defaults, and loss of clients using their credit scoring services.

1. Our process framework based on our data science experience.
   1. Import data
   2. Preprocess
      1. check for missing data
      2. make sure our data types match what our tools require
      3. review & transform data
         1. i.e. treating negative values as neutral or as zero
   3. Exploratory Data Analysis (EDA)
      1. Profile data visualization
      2. Correlation visualization
   4. Feature Selection – Determine the functional form of the model
   5. Modeling
      1. Split the data into training and validation
   6. Evaluation
      1. Assess the model performance
      2. Evaluate
         * 1. the overall fit of the model
           2. the significance of each predictor
           3. the relationship between the target variable and each predictor
   7. Deployment
      1. Deploy the model for real-time prediction
   8. Monitoring
      1. Monitor the predictability of the model over time and rebuild if necessary.
      2. The business environment may change or additional historical data may be provided
2. Descriptions and location of related data sources
   1. Use the raw credit card client historical data provided by Credit One that includes approximately 29,000 customer records with customer demographics (age, sex, relationship status, and education), credit limit, six month credit standing, six month bill amount, six month payment history, and credit status (default or good standing).
3. Managing the Data
   1. Original raw file and cleaned version will be uniquely saved following client data retention and security policies.
   2. We will import and manage data in a Python data analysis library known as Pandas data frame.
4. Any known issues with the data and how you plan to address them
   1. Historical customer data includes an ID field that may skew correlation and modelling studies and should be removed.
   2. The provided historical customer data consists of data types that are both objects and integers.
      1. Transform the object values to integer values.
   3. Negative values in credit standing fields are not a negative factor. They should be treated as a neutral factor.
      1. Transform the negative values to zero values.
5. A flowchart visualizing our detailed process that will be followed, annotated with any potential pitfalls identified and proposed solutions to those pitfalls.
   1. Used a flowchart in section 2 or slide 3 in PowerPoint presentation.
6. Any initial insights you can glean from your quick look at the data.

Program 2 - Task 2:

Now that you have established your process, you're ready to begin your work by preparing and exploring the data. Before we dive in lets review some notes about the project:

 Problem:

1. Increase in customer default rates - This is bad for Credit One since we approve the customers for loans in the first place.
2. Revenue and customer loss for clients and, eventually, loss of clients for Credit One

Investigative Questions:

1. How do you ensure that customers can/will pay their loans? Can we do this?

As you progress through the tasks at hand begin thinking about how to solve this problem. Here are some lessons we learned form a similar problem we addressed last year:

1. We cannot control customer spending habits
2. We cannot always go from what we find in our analysis to the underlying "why"
3. We must on the problem(s) we can solve: What attributes in the data can we deem to be statistically significant to the problem at hand?
4. What concrete information can we derive from the data we have?
5. What proven methods can we use to uncover more information and why?

Random did a relatively great job at correctly predicting Non- Default customers. In conclusion, we cannot control customer spending habits, nor can we change customers’ spending or payment habits. However, we can apply predictive analytics using the customer demographics and historical payment & default data to predict a customer’s likelihood to make payments or default their credit with reasonable accuracy.

The process we used in this task can be applied to existing and new customer base to determine if a person should be extended credit. Based on our findings, we can reduce loan limits to existing customers that are at risk of default. We can decide credit worthiness of new customers using the same features outlined in the EDA dataset.