Define the goal The first step in a data science process is to define a measurable and quantifiable goal.

The stakeholders want to do this project so that they can gain a better understanding of why they are seeing an increase in the number of customers who have defaulted on loans, which could risk them losing business if the problem is not solved right away.

They need to see if they can understand or identify any patterns or trends that will help classify current customers as having a higher risk of defaulting on their debts.

The current solution is not using advanced data science techniques to predict which customers will default based on a number of different features.

The resources that will be needed for this project include using Python and a number of data science packages that are available for Python.

The results of this project will be deployed using AWS and Python.

Collect and manage data This step includes identifying the data you need, then exploring and conditioning it. This is often the most time consuming step.

The data that is currently available is a set of customers and information about past balances and whether or not the customer ended up defaulting on their payments next month or not.

It will hopefully help solve the problem by being able to analyze current customers and their default behavior and use this to predict future customers as having the potential to default or not.

The data quality appears to be solid for now.

Build the model Here is where you try to extract useful insights from the data in order to achieve your goals.

Some of the techniques that might be applied to building the model include analyzing the current set of customer data to get a feel for how the data is structured as well as feature engineering and feature selection to determine which are the most relevant features in predicting future default behavior.

The number of techniques that should be applied will depend on how accurate the models are behaving.

Evaluate and critique the model Once you have derived a model, you need to determine whether it meets your goals. If not, it’s time to loop back to the modeling step.

This model needs to be accurate enough that it exceeds the current thresholds that the stakeholders are using to evaluate current credit default risks.

If the results do not make sense in the context of the financial domain, then we will have to reevaluate and understand why this would be happening.

Present results and document Once you have a model that meets your criteria, you will present your results to your project sponsor and other stakeholders.

The stakeholders should interpret the model as being a tool to help more accurately predict future customer’s default risk, but it should not be communicated that this model is 100% accurate.

It is only as accurate as the previous data that it had to evaluate default risk behavior, but they should feel reasonably confident that it is more accurate than the existing techniques.

They should potentially overrule the model’s predictions when they have information that is not being included as valuable to the model, but they believe could affect whether or not the customer will default.

Deploy and maintain the model Finally the model is put into production but you still need to ensure that the model will run smoothly. In many cases this requires enhancement of the requirements based on customer feedback or in some cases fixing bugs.

The model will be deployed on an AWS server that the stakeholders have access to, so they can manipulate and move the model if they need.

This model should be revised at a set time interval to make sure that it is still producing sufficiently accurate results and there should be thresholds set that require more immediate revision of the model if it drops below the predetermined threshold for accuracy.