**Digital Transformation in Banking Sector Using Data Science With Python**

Summer Internship Report

submitted in partial fulfilment of the requirement for the degree of Bachelor of Computer Application by

**Vansh Kumar Thakur**

**Enrolment No.- 14317702020**

**For Year: 2020- 2023**

****

**Vivekananda Institute of Professional Studies**

(Affiliated to Guru Gobind Singh Indraprastha University)

AU-Block (Outer Ring Road), Pitampura, New Delhi-110034

****

# STUDENT CERTIFICATE

## DECLARATION BY THE CANDIDATE

I the undersigned solemnly declare that the project report **“Digital Transformation in Banking Sector Using Data Science With Python”** is based on my own work carried out during the course of my internship at Ducat Pitampura. I assert the statements made and conclusions drawn are an outcome of my own work. I further certify that

1. The work contained in the report is original and has been done by me under the general supervision of my supervisor.
2. The work has not been submitted to any other Institution for any other degree/diploma/certificate in this university or any other University of India or abroad.
3. The guidelines provided by the university have been kept in mind while writing the report.
4. Reference taken from various sources are given due credit to them in the reference section.

Vansh Kumar Thakur

Enrolment No.- 14317702020

INTERNSHIP CERTIFICATE



# ACKNOWLEDGEMENT

The internship opportunity I had with **Ducat** was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it. I am also grateful for having a chance to meet so many wonderful people and professionals who led me though this internship period.

Bearing in mind previous I am using this opportunity to express my deepest gratitude and special thanks to the MD of **Ducat** who in spite of being extraordinarily busy with her/his duties, took time out to hear, guide and keep me on the correct path and allowing me to carry out my project at their esteemed organization and extending during the training.

I express my deepest thanks to “**Mr. Pushpander Sharma”**, “**Mentor and instructor at Ducat”** for taking part in useful decision & giving necessary advices and guidance and arranged all facilities to make life easier. I choose this moment to acknowledge his/her contribution gratefully

Last but not least, I pay my sincere thanks and gratitude to my parents who always inspire me and keep me motivated.

ABSTRACT

Data science encompasses a set of principles, problem definitions, algorithms, and processes for extracting nonobvious and useful patterns from large data sets. Many of the elements of data science have been developed in related fields such as machine learning and data mining. In fact, the terms data science, machine learning, and data mining are often used interchangeably.

The commonality across these disciplines is a focus on improving decision making through the analysis of data. However, although data science borrows from these other fields, it is broader in scope. Machine learning (ML) focuses on the design and evaluation of algorithms for extracting patterns from data. Data mining generally deals with the analysis of structured data and often implies an emphasis on commercial applications.

Data science takes all of these considerations into account but also takes up other challenges, such as the capturing, cleaning, and transforming of unstructured social media and web data; the use of big-data technologies to store and process big, unstructured data sets; and questions related to data ethics an

# Table Of Contents

Title page I

[Student Certificate II](file:///C:\Users\Vansh\Downloads\harsh%20report%201.docx#_TOC_250008)

[Internship Certificate III](file:///C:\Users\Vansh\Downloads\harsh%20report%201.docx#_TOC_250007)

[Acknowledgement IV](file:///C:\Users\Vansh\Downloads\harsh%20report%201.docx#_TOC_250006)

[Abstract V](file:///C:\Users\Vansh\Downloads\harsh%20report%201.docx#_TOC_250005)

Main Text-

[Chapter 1 - Introduction](file:///C:\Users\Vansh\Downloads\harsh%20report%201.docx#_TOC_250003) 7-10

[**1.1** – Need & Objective……………………………….](file:///C:\Users\Vansh\Downloads\harsh%20report%201.docx#_TOC_250002).....

**1.2** – Methodology………………………………………..

**1.3** – Software Used………………………………………

**1.4** – About Organization…………………………………

[Chapter 2 – Implementation………………………………..1](file:///C:\Users\Vansh\Downloads\harsh%20report%201.docx#_TOC_250001)1-13

[Chapter 3 – Result & Discussion…………………….. 14-](file:///C:\Users\Vansh\Downloads\harsh%20report%201.docx#_TOC_250000)15

**Chapter 4** – Future Scope & Conclusion **16**

**Chapter 5** – **References** ……………….………………………........**17**

**CHAPTER-1**

***INTRODUCTION***

* 1. Need and Objective:

Although banks do many things, their primary role is to take in funds—called deposits—from those with money, pool them, and lend them to those who need funds. Banks are intermediaries between depositors (who lend money to the bank) and borrowers (to whom the bank lends money). The amount banks pay for deposits and the income they receive on their loans are both called interest.

Depositors can be individuals and households, financial and nonfinancial firms, or national and local governments. Borrowers are, well, the same. Deposits can be available on demand (a checking account, for example) or with some restrictions (such as savings and time deposits).

“Digital Transformation in Banking Sector Using Data Science With Python” main motive is to grow the customer base of asset customers (depositors). It is in the best interest of the banks because it will help them to gain more borrowers which will lead to more business via loan interest.

* 1. Methodology:

In Data Science we work on the data which is either provided or collected. The data is of 2 types: Regression type data or Categorical type data. Then that data is cleaned with some tools like pandas, numpy, etc. then the data is analysed for multiple correlations and uniqueness of data. Then after the analysis is performed, a specific model is used based on the problem we have in front of us. After choosing the preferred model, it a trained and tested. Then its accuracy and reliability is checkedwith some more tools like metrices, F1 score, etc. Then after all these processes the program is published.

* 1. Software Used:

The software used in this project is Jupyter notebook. The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at [Project Jupyter](http://jupyter.org/).

Jupyter Notebooks are a spin-off project from the IPython project, which used to have an IPython Notebook project itself. The name, Jupyter, comes from the core supported programming languages that it supports: Julia, Python, and R. Jupyter ships with the IPython kernel, which allows you to write your programs in Python, but there are currently over 100 other kernels that you can also use.

**Installation Process:**

The Jupyter Notebook is not included with Python, so if you want to try it out, you will need to install Jupyter.

There are many distributions of the Python language. This article will focus on just two of them for the purposes of installing Jupyter Notebook. The most popular is [CPython](https://realpython.com/cpython-source-code-guide/), which is the reference version of Python that you can get from their [website](https://www.python.org/). It is also assumed that you are using **Python 3**.

If so, then you can use a handy tool that comes with Python called **pip** to install Jupyter Notebook like this:

$ pip install jupyter

The next most popular distribution of Python is [Anaconda](https://www.anaconda.com/). Anaconda has its own installer tool called **conda** that you could use for installing a third-party package. However, Anaconda comes with many scientific libraries preinstalled, including the Jupyter Notebook, so you don’t actually need to do anything other than install Anaconda itself.

### **Starting the Jupyter Notebook Server**

Now that you have Jupyter installed, let’s learn how to use it. To get started, all you need to do is open up your terminal application and go to a folder of your choice. I recommend using something like your Documents folder to start out with and create a subfolder there called Notebooks or something else that is easy to remember.

Then just go to that location in your terminal and run the following command:

$ jupyter notebook

This will start up Jupyter and your default browser should start (or open a new tab) to the following URL: <http://localhost:8888/tree>

* 1. About Organization:

#### About DUCAT Creative

Getting a job is as difficult as beating the crowd because being in the corporate world demands a lot from the applicant because of which the applicants are putting their best, which results in the increment of difficulty level. You can see each and every thing is connected but the solution of this problem is either spending years to reach to a desired position or come to Ducat. At Ducat we provide the entire necessary computer training which helps the newbies and also the experienced workers so that they can achieve better recognition in this competitive world.

#### How Ducat is Helpful?

Like other educational and training industry at Ducat you will be offered varieties programs but the instructors makes the difference and make Ducat stand out from others. We have a variety of skilled and trained trainers whose approach is different which you can see anywhere. Ducat contributes a lot to the knowledge of its trainees and we try our level hard to contribute the best to increase our trainee’s ability so that they stand out from others and whatever they contribute to the corporate world automatically becomes productive. Not only the fresher but also the corporates who are not able to deal with the rising technology and software are also helped here. We try our level best to deliver our services to every corner of the world by the help of customized education. Our motto is to deliver the best services to you and that is why we have taken the customized approach because we do not want you to compromise with your education.

It is not necessary that you have to leave your job in order to make-up with us. You can contact our experts and can get the best result. To serve you we are always at your service, you can contact us as and when you get time and clear your queries.

#### What types of services are offered by Ducat?

Ducat provides the best available programs which helps in enhancing the technical skills which seems to be beneficial for all the applicants.

* **Software Development:** We provide the best and latest IT software training which helps all the fresher and the corporates to understand well and give them the knowledge to go hand in hand with the latest technologies. This does not only helps the companies but also increases the self-level to deal with all the necessary software.
* **Instructor led campus:** Ducat helps all the new instructors to get the best exposure to show their talent in right way.
* **Workshops and Placement Service:** At Ducat, workshops are held to increase the understanding level because theoretical values are always not enough and workshops helps in getting the practical knowledge which results in better understanding. As everything leads to the placement because if the institute does not provide placement services then it is ultimately bad for the applicants but we provide the best placement services and for that we give our best to give you the best.

**CHAPTER-2**

***IMPLIMENTATION***

Data science is the process of collecting, cleaning, analyzing, visualizing and communicating data to solve problems in the real world.

The implementation of data science in different areas of a company seeks to improve its processes and increase its value. The idea is to transform data into information to increased revenue, reduced costs, increase the business agility and also enhanced customer experience.

Data science something sounds like the holy grail however require extraordinary efforts until to obtain the successful implementation to collect data, process and explore and transform to obtain KPIs, metrics and valuable (and reliable) insights about the business, and even more effort to develop the predictive analysis.

The final objective of data science is to obtain business-focused insights from data. To achieve this, it is vitally important not only the technical implementation but the participation and the involvement of the business areas, in a recurrent process understanding and identifying business opportunities.

**Data Description:**

The dataset has 2 CSV files,

• Data1 - 5000 rows and 8 columns

• Data 2 - 5000 rows and 7 columns

The data consists of the following attributes:

1. ID: Customer ID

2. Age Customer’s approximate age.

3. CustomerSince: Customer of the bank since. [unit is masked]

4. HighestSpend: Customer’s highest spend so far in one transaction. [unit is masked]

5. ZipCode: Customer’s zip code.

6. HiddenScore: A score associated to the customer which is masked by the bank as an IP.

7. MonthlyAverageSpend: Customer’s monthly average spend so far. [unit is masked]

8. Level: A level associated to the customer which is masked by the bank as an IP.

9. Mortgage: Customer’s mortgage. [unit is masked]

10.Security: Customer’s security asset with the bank. [unit is masked]

11.FixedDepositAccount: Customer’s fixed deposit account with the bank. [unit is masked]

12.InternetBanking: if the customer uses internet banking.

13.CreditCard: if the customer uses bank’s credit card.

14.LoanOnCard: if the customer has a loan on credit card

**Aim**

Build a machine learning model to perform focused digital marketing by predicting the potential customers who will convert from liability customers to asset customers.

**Tech stack**

¬ Language - Python

¬ Libraries – numpy, pandas, matplotlib, seaborn, sklearn, pickle, imblearn

**Approach**

1. Importing the required libraries and reading the dataset.

♣ Merging of the two datasets

♣ Understanding the dataset

2. Exploratory Data Analysis (EDA) –

♣ Data Visualization

3. Feature Engineering

♣ Dropping of unwanted columns

♣ Removal of null values

♣ Checking for multi-collinearity and removal of highly correlated features

4 Model Building

♣ Performing train test split

♣ Logistic Regression Model

♣ Weighted Logistic Regression Model

♣ Naive Bayes Model ♣

Support Vector Machine Model

♣ Decision Tree Classifier

♣ Random Forest Classifier

5. Model Validation

♣ Accuracy score

♣ Confusion matrix

♣ Area Under Curve (AUC)

♣ Recall score

♣ Precision score

♣ F1-score

6. Handling the unbalanced data using imblearn.

7. Hyperparameter Tuning (GridSearchCV)

♣ For Support Vector Machine Model

8. Creating the final model and making predictions

9. Save the model with the highest accuracy in the form of a pickle file.

**CHAPTER-3**

***RESULT AND DISCUSSION***

1. Understanding the business problem.

2. Importing the dataset and required libraries:

1. Pandas

2. Numpy

3. Matplotlib

4. sklearn train\_test\_split

5. sklearn linear Model

6. sklearn DecisionTreeClassifier

3. Performing basic Exploratory Data Analysis (EDA):

1. column names were simplified

2. setting the index

4. Removal of unwanted features and missing data handling if required, using appropriate methods:

1. Null values were removed using fillna.

2. merging both the data sets using the hyper parameter outer

5. Checking data distribution using statistical techniques.

6. Using python libraries such as matplotlib and seaborn for data interpretation and advanced visualizations.

7. Splitting Dataset into Train and Test using sklearn.

8. Training a model using Classification techniques like Logistics Regression, Naïve Bayes, Decision Tree Classifier, Random Forest Classifier and Support Vector Machine.

Decision Tree Classifier is used.

9. Tuning hyper-parameters of models to achieve optimal performance.

10. Making predictions using the trained model

Prediction made by using decision tree classifier

11. Gaining confidence in the model using metrics such as accuracy score, confusion matrix, recall, precision and f1 score

Accuracy and F1 score is used

12. Handling the unbalanced data using various methods.

No unbalance in the data found

13. How Target variable is dependent on the values of Input features.

14. Selection of the best model based on performance metrics:

Decision Tree Classifier is used as it showed the best accuracy score

15. Saving the best model in pickle format for future use.

CHAPTER-4

***FUTURE SCOPE AND CONCLUSIONS***

This model can be used by the organizations to improve there productivity and increasing their business. It will help the organizations to focus on the potential customers instead of focusing on all the customers.

As we can see that the usage of data science is increasing day by day in every field of work. And in future there will be much more development in this software then there is at the moment.

Future works that are needed to be done before publishing this software are:

1. Connecting it to the database
2. Gathering more data to make the software more reliable
3. Building up a website so that its services can be provided globally

***REFERENCES:***

***References taken for Jupyter Notebook:*** [***www.realpython.com***](http://www.realpython.com)

***References taken for data science explanation:***

[***www.thinkingondata.com***](http://www.thinkingondata.com)

***References taken for Banking problems:***

[***www.imf.org***](http://www.imf.org)

***A huge helping hand of our mentor in this who training program Mr. Pushpander Sharma who was there to solve all the doubts that occurred in my mind.***

***At end thanks allot to VIPS faculty who provided us this opportunity to explore other skills and find their best route to move on in our lives.***