

# **Summer Internship Report**

Submitted

by

**Tatikonda Lakshman - AP19110010385**

**Kamisetty Smaran - AP19110010350**

**Dandibhotla Bala Pranav - AP19110010428**

**Katakam Akshay - AP19110010425**

**Batchu Raja Nithin- AP19110010495**

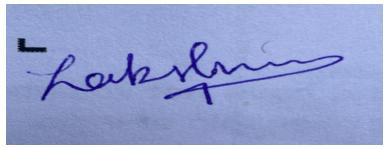


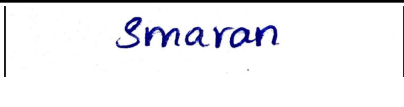

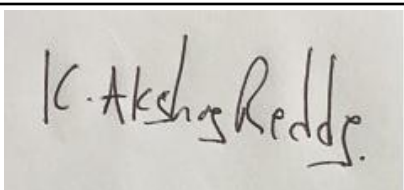
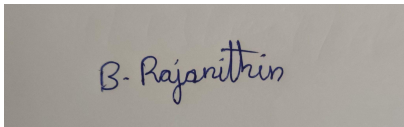
**Department of Computer Science and Engineering**

**SRM University-AP, Andhra Pradesh, India**

**July - 2021**

## **DATASHEET**

Roll Numbers	:	1. AP19110010385 2. AP19110010350 3. AP19110010428 4. AP19110010425 5. AP19110010495
Names of the student	:	1. Tatikonda Lakshman 2. Kamisetty Smaran 3. Dandibhotla Bala Pranav 4. Katakam Akshay 5. Batchu Raja Nithin
Branch & Section	:	CSE C
Batch	:	2019-2023
Type of internship	:	Industry internship
Company Name/Institute Name	:	Andhra Pradesh State Skill Development Corporation (APSSDC)
Company/Institute Website	:	<a href="https://www.apssdc.in/home/homepage">https://www.apssdc.in/home/homepage</a>
Start Date (MM/DD/YYYY)	:	06/07/2021
End Date (MM/DD/YYYY)	:	07/15/2021
Duration (No. of days)	:	39
Status of the internship	:	Completed
Name of internship mentor (SRM Faculty)	:	Dr Asish Bera (SRM Faculty)
Names & Signatures of the student	:	Tatikonda Lakshman 

		Kamisetty Smaran	
		Dandibhotla Bala Pranav	
		Katakam Akshay	
		Batchu Raja Nithin	

### **ACKNOWLEDGEMENT**

The internship opportunity we had with Andhra Pradesh State Skill Development Cooperation (APSSDC) was a great chance for learning and applying my learnings in various tasks. We are very grateful for the opportunity provided to be a part of this internship. We also acknowledge our gratitude to all those who have helped put our ideas to perfection and have assigned tasks, well above the level of simplicity and into something concrete and unique.

We, wholeheartedly thank Polamarasetti Lavanya, Golla Naga Mounika, Aitam Sri Sarojini Niharika for having faith in us, and explaining each topic with great clarity and taking time to clear all our doubts.

We use this opportunity to express our deepest gratitude and special thanks to our university (SRM AP) Corporate Relations & Career Services for providing us with this online internship opportunity in these uncertain times of covid crisis.

We perceive this opportunity as a big milestone in our career development. We will strive to use gained skills and knowledge in the best possible way, and we will continue to work on their improvement, to attain desired career objectives. Hope to continue cooperation with all of you in the future.

## **TABLE OF CONTENTS**

<b>Sl. No.</b>	<b>Content</b>	<b>Page No.</b>
1	<b>Introduction</b>	5
2	<b>Objective of the Internship</b>	5
3	<b>Skills acquired through the internship</b>	5
4	<b>Overview of the project carried out during internship</b>	6
5	<b>Results</b>	7
6	<b>Conclusion</b>	8
7	<b>References (If any, otherwise remove this)</b>	9

## I. INTRODUCTION:

This course will introduce data manipulation and cleaning techniques with python pandas data science library and Data Frame as the central data structures for data analysis. NumPy provides the most fundamental module for scientific computing with python. It also supports multidimensional arrays. Pandas is the most popular Python library for manipulating data. Pandas is an extension of NumPy. The underlying code for Pandas uses the NumPy library extensively. The primary data structure in Pandas is called a data frame.

## II. OBJECTIVE OF THE INTERNSHIP:

Data scientists are responsible for discovering insights from massive amounts of structured and unstructured data to help shape or meet specific business needs and goals.

The key objective of data science is to extract valuable information for use in product development, decision making and forecasting.

These insights and reports help companies analyse their marketing strategies, make powerful data-driven decisions and create better advertisements.

## III. SKILLS ACQUIRED THROUGH THE INTERNSHIP:

**Data Scraping** - Gathering data from websites is one of the most logical and easily accessible sources of data.

**Data Frames** - SQL is important in data science and great for handling large amounts of data however it lacks Machine Learning and Data Visualization.

**Data Visualization** - Data science is about communicating your findings and data visualization is an incredibly valuable part of that.

**Machine learning** - A lot of data science can be done with select, join and group. But sometimes you need to do some non-trivial machine learning.

**Python Programming** - Python is a general-purpose programming language having multiple data science libraries.

**Statistics** - Machine Learning starts out as statistics and then advances. The knowledge of the concept of descriptive statistics like mean, median, mode, variance, the standard deviation is a must.

#### IV. OVERVIEW OF THE PROJECT/WORK CARRIED OUT DURING INTERNSHIP:

##### **Tasks:**

We have completed many tasks throughout the internship. Tasks based on Data types(strings, modules, lists, Dictionaries, tuples), loops, arrays, functions, map, operating system services, recursion, sorting, binary search tree, python NumPy, python pandas.

##### **Project:**

Project Github Link:

<https://github.com/raiperiod2002/AP19110010495-Internship/blob/main/Internship%20Group%206%20Project%20.%20%20.ipynb>

##### **Our Approach:**

Cashless payments are the go-to method of payment in this day and age. So, it is important for credit card companies to recognize any fraudulent credit card transactions so that customers are not charged for extra items.

##### **Context:**

Data set source - <https://www.kaggle.com/mlg-ulb/creditcardfraud>

The datasets contain transactions made by credit card European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. It only contains numerical variables resulting from a PCA transformation. Features V1, V2, ... V28 are the principal components obtained with PCA. It has 'Time' and 'Amount' features. 'Time' contains the seconds elapsed between each transaction and the first transaction. The feature 'Amount' represents the transaction amount. Feature 'Class' is the response variable and it gives value '1' in case of fraud and '0' if the transaction is legitimate.

##### **Data set analysis:**

We performed Dataset exploration to find out the following information

- How large is the imbalance b/w features and variables
- imbalance to check if its a fraud payment
- Correlation b/w features and variables
- Is there a time period when the frauds high
- SMOTE to balance feature in both sets

When we compared the number of real transactions with the number of fraud transactions in the data set we found that the dataset is highly imbalanced which can significantly affect our ML model. A balanced dataset is best for training purposes

We also found that there is no pattern for when fraudulent activities occur. They happen at random. The Transaction amount for each fraudulent transaction is significantly lower compared to the Real transactions. This makes it a lot more difficult to detect fraudulent activities. The max amount is approximately \$2100.

We observed the mean and standard deviation for the VI-V28 feature and visualized it.

We also observed the correlation Matrix to see how each variable is related to the other and the target variable.

## **V. RESULTS/OUTPUT:**

### **Tasks:**

Throughout the internship, we have uploaded several tasks to GitHub and our internship trainers have reviewed our progress. GitHub Task Submissions Link :

Tatikonda Lakshman - <https://github.com/lakshman666/APSSDC-TASKS>

Kamisetty Smaran - <https://github.com/SmaranK/AP19110010350-Smaran>

Dandibhotla Bala Pranav - <https://github.com/pranav1028/Python-task--AP19110010428---D-Bala-Pranav>

Katakam Akshay - <https://github.com/akshay-rdy/AP19110010425-Akshay>

Batchu Raja Nithin - <https://github.com/rain2002/AP19110010495-B-Raja-Nithin>

### **Project:**

#### **Observations:**

- Isolation Forest detected 73 errors versus Local Outlier Factor detecting 97 errors vs. SVM detecting 8516 errors
- Isolation Forest has a 99.74% more accurate than LOF of 99.65% and SVM of 70.09
- When comparing error precision & recall for 3 models, the Isolation Forest performed much better than the LOF as we can see that the detection of fraud cases is around 27 % versus the LOF detection rate of just 2 % and SVM of 0%.

- So overall Isolation Forest Method performed much better in determining the fraud cases which is around 30%.
- StandardScaler removes the mean and scales each feature/variable to unit variance. This operation is performed feature-wise in an independent way. StandardScaler can be influenced by outliers (if they exist in the dataset) since it involves the estimation of the empirical mean and standard deviation of each feature.
- RobustScaler transforms the feature vector by subtracting the median and then dividing by the interquartile range (75% value — 25% value). Like MinMaxScaler, our feature with large values — normal-big — is now of similar scale to the other features.
- Logistic regression has a macro average of f1 score is 0.87 whereas sGD has 0.82, Decision tree has 0.87 and RFC has 0.94.
- So Random forest classifier is the best in these 4 ways to find fraud detection.
- We can also improve on this accuracy by increasing the sample size or use deep learning algorithms however at the cost of computational expense. We can also use complex anomaly detection models to get better accuracy in determining more fraudulent cases

## **VI. CONCLUSION:**

Overall, we would describe our internship as a positive and instructive experience. This Internship has provided a great insight into data science that we never thought of until we started this internship. At the end of this internship, we are very confident in using the skills we have gained here for our future work.

We have greatly improved our statistical approach towards data science from this internship. We can conclude by saying this internship has been an excellent and rewarding experience. We were able to gain practical knowledge. We could not be more grateful.



## VII. REFERENCES (If Any)

[1] L.J.P. van der Maaten and G.E. Hinton, [Visualizing High-Dimensional Data Using t-SNE](#) (2014), Journal of Machine Learning Research

[2] Machine Learning Group — ULB, [Credit Card Fraud Detection](#) (2018), Kaggle

[3] Nathalie Japkowicz, [Learning from Imbalanced Data Sets: A Comparison of Various Strategies](#) (2000), AAAI Technical Report WS-00-05