REPORT

ON

FOUR WEEKS OF INTERNSHIP
Carried out at

INMOVIDU TECHNOLOGIES

Submitted to

NMAM INSTITUTE OF TECHNOLOGY, NITTE
(An Autonomous Institution under VTU, Belagavi)

In partial fulfillment of the requirements for the award of the

Degree of Bachelor of Engineering
in
Electronics and Communication Engineering

by

Varun N Naik
USN **4NM17EC172**

Under the guidance of

Mr. Vignesh Shenoy

(Assistant Vice President, InMovidu Technologies Pvt. Ltd)

NITTE EDUCATION TRUST | N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India

NITTE EDUCATION TRUST | N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India

CERTIFICATE

This is to certify that the "Internship report" submitted by Mr. Varun N Naik bearing USN 4NM17EC172 *of 8th semester B.E., a bonafide student of NMAM Institute of Technology, Nitte, has undergone eight weeks of internship at Inmovidu Technologies Pvt. Ltd during May 2020 fulfilling the partial requirements for the award of degree of Bachelor of Engineering in* ***Electronics and Communication Engineering*** *at NMAM Institute of Technology, Nitte.*

Shubha B.

Name and Signature of Mentor

14/7/21

Signature of HOD

INDUSTRY CERTIFICATE

ACKNOWLEDGEMENT

This project would not have been successful without the encouragement, guidance and support of various personalities. First and foremost, I would like to express my sincere gratitude towards my mentor, **Mr. Ankur Khanna** and **Mr. Vignesh Shenoy**, Assistant Vice President for their guidance, encouragement and inspiration throughout the work.

I extend my gratitude to **Dr Rekha Bhandarkar**, Professor and Head, Department of Electronics and Communication Engineering, N. M. A. M. Institute of Technology, Nitte, for her encouragement and providing necessary facilities in the department. I wish to acknowledge the support of **Dr Niranjan N. Chiplunkar**, Principal, N. M. A. M. Institute of Technology, Nitte, for providing a motivational academic environment.

I wish to express my sincere thanks to Inmovidu Technologies for providing an opportunity to carry out the project at their esteemed institution. Many thanks to all the teaching and non-teaching staff of the Department of Electronics and Communication Engineering, N. M. A. M. Institute of Technology, Nitte, for their patience and help during my project work.

Finally, I would like to thank all my friends and well-wishers who have helped me whenever needed and supported me in the completion of the work.

Varun N Naik

Table of Contents

Abstract

Inmovidu provides an online simulation of the classroom environment with a focus on comprehensive teaching and training on current tech-industry specializations that are constituents of the ‘Industry 4.0’ principle like the “Internet of Things, Cyber Security, Artificial Intelligence” among others.

The internship at Inmovidu Technologies was conducted for a total of 8 weeks under the guidance of a mentor, with the first half consisting of training and the second half consisting of the project. The internship started with the basics of Python and machine learning. Moreover, basic concepts of statistics were taught which were necessary to understand the data. There were several Python libraries used in machine learning such as NumPy, Pandas, Scikit used for Exploratory Data Analysis (EDA). Finally, different models such as Linear Regression, Logistic Regression, Classification and Clustering were utilized and the results were compared.

The project consisted of two parts, the EDA of the data and the modelling of the data. With the help of this internship, the theoretical knowledge was transformed into hands-on practical skills. Machine learning is one of the fast-growing technologies which has wide applications ranging from business to e-healthcare. This internship has given a glimpse into how organizations work on real-world projects by increasingly relying on machine learning to solve problems. There are many benefits to learning machine learning ranging from performing automation without the intervention of humans to identifying trends and patterns. This internship was imperative as machine learning and data science become increasingly used in the future.

Introduction of the Industry / Company

Inmovidu stands for Internet Mobile Virtual Education, located in Bangalore, India with around 50 employees, which offers certification based industrial training and internship opportunities to undergraduate students. It also provides an online simulation of the classroom environment with a focus on comprehensive teaching and training on current tech-industry specializations that are constituents of the 'Industry 4.0' principle like the "Internet of Things, Cyber Security, Artificial Intelligence" among others. The students undergo rigorous training in the tenets relating to the domain of their choice and work on projects that simulate real-life situations for them to gain invaluable experience while working from the comforts of their environment.

Inmovidu aims to make people self-reliant in every sphere of life by imparting them the technical and tactical know-how of the concepts of self-education ranging from technical domains like Cyber Security, Artificial Intelligence, Machine Learning to everyday deeds like Fitness & Health, Cooking, Personality Development. It is a professional community of Industry Experts and academia, who have come together to help learners become employable with an emphasis on design thinking and a learner-centred approach to problem-solving and focusing on application-based learning. It aims at specialization into industrial training, live interactive session, technical certification, top-notch industry experts for training and career counselling. Inmovidu is the authorized partner for extending Microsoft Technology Associate (MTA), Autodesk certified user and Adobe-certified associates.

.

Details of the training undergone

I. Training

To understand machine learning tools, the internship dived into the basics of Python programming, an approachable and well-known programming language. This provided the foundation of concepts for data analysis and visualization, which would be vital for the following project. This was followed by studying various supporting libraries required for machine learning analysis. Different types of models were explored such as supervised and unsupervised learning. These included popular algorithms such as Classification, Regression, and Clustering and popular models such as Train/Test Split, Root Mean Squared Error (RMSE). Moreover, the models were studied using real examples.

Python is a popular object-oriented programing language having the capabilities of a high-level programming language. Its easy to learn syntax and portability capability makes it popular these days. As being an open-source programming language, Python is supported by a very large developer community. Due to this, the bugs are easily fixed by the Python community. This characteristic makes Python very robust and adaptive. The internship explored basics of Python such as data types, data operators, data structures, control flow statements and loops, reading and writing in Python, comments and indentation, defining function, error handling and modules.

Statistical methods are required in the preparation of train and test data for the machine learning model which includes techniques for outlier detection, missing value imputation, data sampling and data scaling. It is also used when evaluating the skill of a machine learning model on data not seen during training and when selecting a final model or model configuration to use for a predictive modelling problem. It provides methods for summarizing raw observations into information that we can understand and share for instance, to summarize properties of the sample of data, such as the common expected value (e.g. the mean or median) and the spread of the data (e.g. the variance or standard deviation). Descriptive statistics may also cover graphical methods that can be used to visualize samples of data. Concepts such as measures of central tendency, measures of dispersion, measures of shape, central limit theorem and hypothesis testing were also studied.

NumPy is the core library for scientific computing in Python. It provides a high-performance multidimensional array object, and tools for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O and much more. Pandas is mainly used for data analysis. Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL, Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.

Data visualization is the technique to present the data in a pictorial or graphical format. It enables stakeholders and decision-makers to analyze data visually. The data in a graphical format allows them to identify new trends and patterns easily. The Matplotlib library has emerged as the main data visualization library. Matplotlib is a Python two-dimensional plotting library for data visualization and creating interactive graphics or plots. Using Python's Matplotlib, the data visualization of large and complex data becomes easy. Seaborn is a Python visualization library based on Matplotlib. It provides a high-level interface to draw attractive statistical graphics.

Anaconda comes with pre-built libraries that save a lot of time for projects like ML to get started quickly. The exercises were performed on Anaconda using tools such as Jupyter Notebook. The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. It can be used for data cleaning and transformation, data visualization, machine learning, and much more.

Exploratory Data Analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. Various machine learning algorithms were also looked at. All machine learning models are categorized as either supervised or unsupervised. Supervised learning involves learning a function that maps an input to an output based on example input-output pairs. Examples of supervised learning are Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression etc. Unlike supervised learning, unsupervised learning is used to draw inferences and find patterns from input data without references to

labelled outcomes. Examples of Unsupervised Learning are the Apriori algorithm and K-means.

II. **Project**:

Problem statement: On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic, sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone on board, resulting in the death of 1502 out of 2224 passengers and crew. Hence, build a predictive model that answers the question: "what sorts of people were more likely to survive?" using passenger data (i.e. name, age, gender, socio-economic class, etc.).

Objective: To perform Data analysis to predict the survival rate of passengers from Titanic using tools of Machine Learning and Python.

Methodology: The data was acquired in a .csv file from the Kaggle website as described in Table 1. The work was divided into two parts, the EDA of the dataset and the analysis of various models. The EDA further consisted of steps such as variable identification, univariate analysis, bivariate analysis, missing values treatment, outlier treatment, variable creation and transformation as shown in Figure 1.

Table 1: Titanic Dataset Details

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

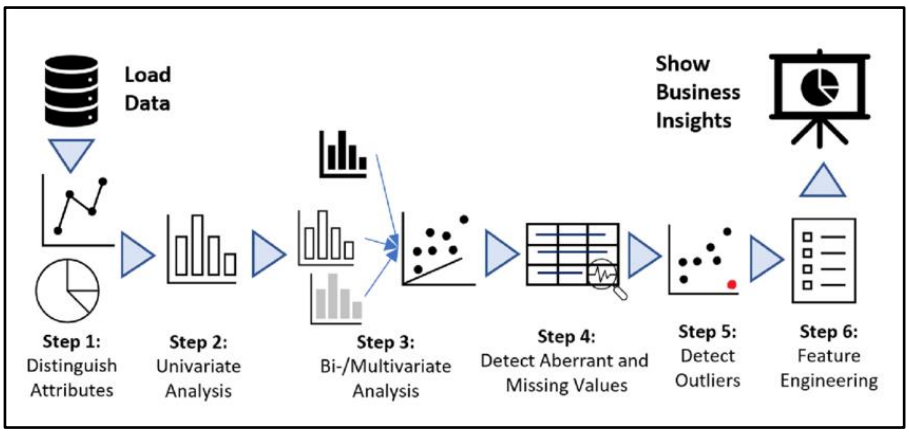

Figure 1: Steps involved in EDA

The first step is to understand the variables and identify their types i.e. Numerical or Categorical. Then, during variable identification, Survived was selected as the dependent variable and Pclass, Sex, Age, SibSp, Parch, Fare and Embarked were classified as independent variables. The next step is cleaning the dataset by removing variables with a high number of null values. Followed by this, it is important to analyze the relationships between variables. Hence, univariate analysis is performed. This can be in the form of a Histogram, count plot or boxplot. Bi-variate Analysis finds out the relationship between two variables. Bi-variate analysis can be performed for any combination of categorical and continuous variables.

Missing data in the training data set can reduce the power/fit of a model or can lead to a biased model if the behaviour and relationship with other variables is not analyzed correctly. It can lead to wrong prediction or classification. The reasons for the occurrence of these missing values may be due to wrong data extraction or data collection. Methods to treat missing values are deletion and Mean/Mode/Median imputation. In this case, the null values of Age are replaced with the mean age and null values of Embarked are replaced by the mode S.

However, it is important to look at the Outliers which are the observations that appear far away and diverge from an overall pattern in a sample as they can drastically change the results of the data analysis and statistical modelling. The most commonly used method to detect outliers is visualization. But, any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR (IQR or InterQuartile Range is the difference between 75th and 25th percentiles) or any value which out of the range of 5th and 95th percentile can be considered as an

outlier and can be capped at those values. In this way, outliers of Age, Fare and Parch variables cab be dealt with. Furthermore, by dummy variable creation it is possible to extract more information from existing data making it more useful. Here, categorical variables like Embarked and Sex are converted into numerical dummy variables. Each process is vital and corrections at each step add up and thus ensure the quality of the prediction model. This processed dataset forms the foundation for building the various models. Major insights extracted were that females were given higher priority in the rescue operation than males, the first-class people were given higher priority than the second class and the third class, the features such as age, Siblings on board and Parents on board didn't have a major influence on the survival probability.

1. Logistic Regression

```
In [182]: from sklearn.linear_model import LogisticRegression
          logReg = LogisticRegression()
          logReg.fit(X_train, y_train)

Out[182]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                             intercept_scaling=1, l1_ratio=None, max_iter=100,
                             multi_class='auto', n_jobs=None, penalty='l2',
                             random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
                             warm_start=False)

In [183]: y_pred = logReg.predict(X_test)
```

Figure 2: Building the Logistic Regression Model

Training of each model begins with defining the X and y variables. Since it is needed to find out how many passengers survived, the values of the Survived column are considered as “y” and the rest of the columns are considered as “X”. Further, the data is separated into test and train datasets done using the Sklearn module’s train_test_split() function. The models are defined using the respective model function and trained using the fit() function as shown in figure 2. The model once built, is used on the test data to predict the y variable. To evaluate the performance of the model, confusion matrix, a classification report and accuracy score can be used. However, these methods do not produce reliable results. Hence, K-fold Cross-Validation (CV) provides a solution to this problem by dividing the data into folds and ensuring that each fold is used as a testing set at some point. The same dataset was used to train various models such as Logistic Regression, K-Nearest Neighbors Classifier (KNN), Decision tree and Random forest with an accuracy score of 0.79, 0.80, 0.77, 0.78 respectively as depicted in Figure 3, 4, 5 and 6. It is observed that the KNN model has the highest accuracy among all the models.

Result:

Logistic Regression

```
scores_logReg = cross_val_score(logReg, X_train, y_train, cv=10, scoring='accuracy')
print(scores_logReg)
print(scores_logReg.mean())
```

```
[0.74603175 0.80952381 0.80952381 0.77419355 0.79032258 0.79032258
 0.77419355 0.83870968 0.82258065 0.80645161]
0.7961853558627752
```

Figure 3: Accuracy Score for Logistic Regression Model

knn

```
scores_knnmodel = cross_val_score(knnmodel, X_train, y_train, cv=10, scoring='accuracy')
print(scores_knnmodel)
print(scores_knnmodel.mean())
```

```
[0.82539683 0.82539683 0.74603175 0.79032258 0.80645161 0.79032258
 0.74193548 0.83870968 0.88709677 0.79032258]
0.8041986687147977
```

Figure 4: Accuracy Score for KNN Model

Decision Tree

```
scores_decTree = cross_val_score(decTree, X_train, y_train, cv=10, scoring='accuracy')
print(scores_decTree)
print(scores_decTree.mean())
```

```
[0.71428571 0.84126984 0.80952381 0.75806452 0.77419355 0.75806452
 0.74193548 0.74193548 0.82258065 0.82258065]
0.7784434203789041
```

Figure 5: Accuracy Score for Decision Tree Model

Random Forest

```
scores_ranFor = cross_val_score(ranFor, X_train, y_train, cv=10, scoring='accuracy')
print(scores_ranFor)
print(scores_ranFor.mean())
```

```
[0.74603175 0.79365079 0.80952381 0.80645161 0.79032258 0.75806452
 0.72580645 0.77419355 0.82258065 0.82258065]
0.7849206349206348
```

Figure 6: Accuracy Score for Random Forest Model

Conclusion

The internship program at InMovidu Technologies conducted over a course of 8 weeks has enabled interns to work on machine learning tools to provide solutions to real-world problems.

Interns were able to get a glimpse into the fundamental concepts ranging from Python and statistics to machine learning tools. It provided the foundation for using Python for data analysis and visualization and explored various libraries available with it which greatly simplify the process. Interns, learnt first-hand how Machine Learning extracts meaningful insights from raw data to quickly solve complex, data-rich business problems and how ML algorithms learn from the data iteratively and allow computers to find different types of hidden insights without being explicitly programmed to do so. Interns also studied the relationship between statistics and machine learning and its need to successfully analyze data. It introduced various types of machine learning algorithms and their specific uses and advantages.

The internship gave an opportunity to the interns to gain valuable work experience, have better work habits, possess excellent soft skills, and have higher technical and industry skills by receiving formal job training and professional guidance. Thus, the guidance of industry experts has enabled interns to learn new techniques and skills in technologies that will be in high demand in the future.

References

1. "Data Visualization in Python using Matplotlib: Simplilearn". [Online]. Available: www.simplilearn.com/data-visualization-in-Python-using-matplotlib-tutorial.
2. Géron, Aurélien. “Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques”. O'reilly Media, 2019.
3. Kubat, Miroslav. “Introduction to Machine Learning”. Springer International pu, 2018.
4. Masis, Serg. “Interpretable Machine Learning with Python”. Packt Publishing Limited, 2021.
5. Müller, Andreas Christian, and Sarah Guido. “Introduction to Machine Learning with Python”. O'reilly Media, 2018.
6. “Titanic - Machine Learning from Disaster.”. [Online]. Available: www.kaggle.com/c/titanic/overview.