**PROJECT REPORT**

**Project Name**

*Submitted towards the partial fulfillment of the criteria for award of KPMG Data Science Prodegree by Imarticus*

*Submitted By:*

*Member 1 (Roll #)*

*Course and Batch: DSP Month Year*



# Abstract

**Keywords**

*Disclaimer: \*Data shared by the customer is confidential and sensitive, it should not be used for any purposes apart from capstone project submission for DSP. The Name and demographic details of the enterprise is kept confidential as per their owners’ request and binding.*

# Acknowledgements

**[To be changed by candidates as per their requirement]**

We are using this opportunity to express my gratitude to everyone who supported us throughout the course of this group project. We are thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, we were fortunate to have \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ as our mentor. He/She has readily shared his immense knowledge in data analytics and guide us in a manner that the outcome resulted in enhancing our data skills.

We wish to thank, all the faculties, as this project utilized knowledge gained from every course that formed the DSP program.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: April 01, 2018 Member 1

Place: Mumbai

# Certificate of Completion

**[To be changed by candidates as per their requirement]**

I hereby certify that the project titled “Title comes here” was undertaken and completed under my supervision by Member 1 and Member 1 from the batch of DSP (Apr 2018)

Mentor: Mentor Name (if any)

Date: April 1, 2018

Place – Mumbai

**Table of Contents [Sample purpose only]**

[Abstract 2](#_heading=h.gjdgxs)

[Acknowledgements 2](#_heading=h.30j0zll)

[Certificate of Completion 3](#_heading=h.3znysh7)

[CHAPTER 1: INTRODUCTION 6](#_heading=h.2et92p0)

[1.1](#_heading=h.tyjcwt) Title & Objective of the study 6

[1.2](#_heading=h.3dy6vkm) Need of the Study 6

[1.3 Business or Enterprise under study 6](#_heading=h.1t3h5sf)

[1.4 Business Model of Enterprise 6](#_heading=h.4d34og8)

[1.4 Data Sources 6](#_heading=h.2s8eyo1)

[1.5 Tools & Techniques 6](#_heading=h.17dp8vu)

[1.6 Infrastructure Challenges 6](#_heading=h.3rdcrjn)

[CHAPTER 2: DATA PREPARATION AND UNDERSTANDING 6](#_heading=h.26in1rg)

[2.1 Phase I – Data Extraction and Cleaning: 6](#_heading=h.lnxbz9)

[2.2 Phase II - Feature Engineering 6](#_heading=h.35nkun2)

[2.3 Data Dictionary: 7](#_heading=h.1ksv4uv)

[2.4 Exploratory Data Analysis: 7](#_heading=h.44sinio)

[CHAPTER 3: FITTING MODELS TO DATA 7](#_heading=h.2jxsxqh)

[4.1 LINEAR REGRESSION MODEL 7](#_heading=h.z337ya)

[4.1.1 First Linear Regression Model 7](#_heading=h.3j2qqm3)

[4.1.2 Second Linear Regression model 7](#_heading=h.1y810tw)

[4.1.3 Third Linear Regression model 7](#_heading=h.4i7ojhp)

[4.2 RANDOM FOREST 8](#_heading=h.2xcytpi)

[4.2.1 INFRASTRUCTURE CHALLENGES 8](#_heading=h.1ci93xb)

[CHAPTER 5: KEY FINDINGS 8](#_heading=h.3whwml4)

[CHAPTER 6: RECOMMENDATIONS AND CONCLUSION 9](#_heading=h.2bn6wsx)

[CHAPTER 7: REFERENCES 9](#_heading=h.qsh70q)

**List of Figures**

**List of Tables**

# CHAPTER 1: INTRODUCTION

## Title & Objective of the study

## Need of the Study

## 1.3 Business or Enterprise under study

## 1.4 Business Model of Enterprise

## 1.4 Data Sources

## 1.5 Tools & Techniques

**Tools:**

**Techniques:**

## 1.6 Infrastructure Challenges

# CHAPTER 2: DATA PREPARATION AND UNDERSTANDING

One of the first steps we engaged in was to outline the sequence of steps that we will be following for our project. Each of these steps are elaborated below

## 2.1 Phase I – Data Extraction and Cleaning:

* **Missing Value Analysis and Treatment**
* **Handling Outliers**
* **Feature Extraction**

## 2.2 Phase II - Feature Engineering

## 2.3 Data Dictionary:

## 2.4 Exploratory Data Analysis:

# CHAPTER 3: FITTING MODELS TO DATA

**[To be changed by candidates as per their requirement]**

**[The below is for example purpose only]**

**Linear regression**

**Random forest**

## 4.1 LINEAR REGRESSION MODEL

## 4.1.1 First Linear Regression Model

## 4.1.2 Second Linear Regression model

**4.1.3 Third Linear Regression model**

**4.1.4 Model Validation**

**4.1.4.1 Checking for Variance inflation (Multicollinearity)**

**4.1.4.1.1 Fourth Linear Regression model**

**4.1.4.1.2 Fifth Linear Regression model** -

**4.1.4.2 Checking for influence points via Cook’s Distance**

**4.1.5 Predicting New Data**

**4.1.6 Comparison of Linear Regression model**

# 4.2 RANDOM FOREST

We applied Random Forest on the Training data set to validate if any further improvement of the model can be performed post the linear regression. Below were the parameters which were applied for Random Forest

## 4.2.1 INFRASTRUCTURE CHALLENGES

**CHAPTER 5: KEY FINDINGS**

Significant Variables identified in linear models are also used in Random forest

Below table provides a snapshot of the various models which the business can choose from based on the pros and cons of each model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sr. No | Model Name | RMSE | Benefits | Trade-offs |
| 1 | Linear\_Regression\_Model | 0.70 | High interpretability  Significant variables can be easily identified | May give lower accuracy |
| 2 | Random\_Forest\_Model | 0.20 | Highest Accuracy | Minimal / lower interpretability |

Below are some of the key findings:

* **Linear Regression:**
* **Random Forest:**

**CHAPTER 6: RECOMMENDATIONS AND CONCLUSION**

**CHAPTER 7: REFERENCES**