 Credit card Default Prediction

DPR

DETAIL PROJECT REPORT

### Q1.What is the Problem statement ?

 **This project is “Credit Card Default Prediction”**

**The goal of this project is to predict the defaulter on the basis of different parameters(or features).**

**First of all, any user is needed to register within our website. After this, user has to pass different values such as gender, Limit Balance, Education, Sex, Previous EMI status, Previous Payment Status, etc.**

**These values provided by the users are stored in the database, and on the backend, we pass this datasets with all the models and then select score with maximum accuracy.**

**So, overall this project is able to predict the probability of whether the user will be default or not.**

**This dataset has 30k rows and 25 features.**

## Q2 : Which tools You’ve used?

 Language : Python

Libraries : Sklearn , pandas , numpy , .git , matplotlib , seaborn Clouds : AZURE (Blob Storage, Databricks, Delta Lake Storage, Azure Pipeline)

Storage : Azure Blob Storage IDE : pycharm / Vscode WebFrame work : Flask

Front end Laguages : HTML , CSS etc

### Q3. What was the data type?

 **In this dataset, we have 25 columns of numeric data type.**

Q4 . **What Hadoop distribution were you using?**

 **In this project, we have used Azure delta lake’s blob storage to store the data and azure databricks to process the code.**

### Q5. What techniques were you using for data pre-processing for various data science use cases and visualization?

 **There are multiple steps that we do for data preprocessing, like data cleaning, data integration, data scaling, etc. Some of them are listed as follows:**

* **For Machine Learning: While preparing data for a model, data should be verified using multiple tables or files to ensure data integrity.**
* **Identifying and removing unnecessary attributes.**
* **Identifying, filling or droping the rows/columns containing missing values based on the requirement.**
* **Identifying and removing outliers**
* **Scaling the data so that the difference between the magnitudes of the data points in different columns are not very big.**

### Q6: What type of problem you will be facing in future with this project ?

 It’s very Good Question …

1. **As we move ahead , we’ll be receiving more and more data ,and to make the**

system more & more Robust we need to train the system on (Old data + new data )

1. **As the Data Volume will high , it’ll take more time to Learn & Costly , at that time requirement of resources will increases ..**
2. **To store the Old data & new data , we’ll be requiring the High storage Capacity ..**
3. **To train the model on the Such Huge dataset , System configuration need to be updated in future ..**
4. **So there will be chance that increase in Handling Charges….**
5. **If we want to train model on new data , we need to replace entire system with**

new system from scratch ( it’ll be Painfull)

### Q7:How you can get rid from Above Problem?

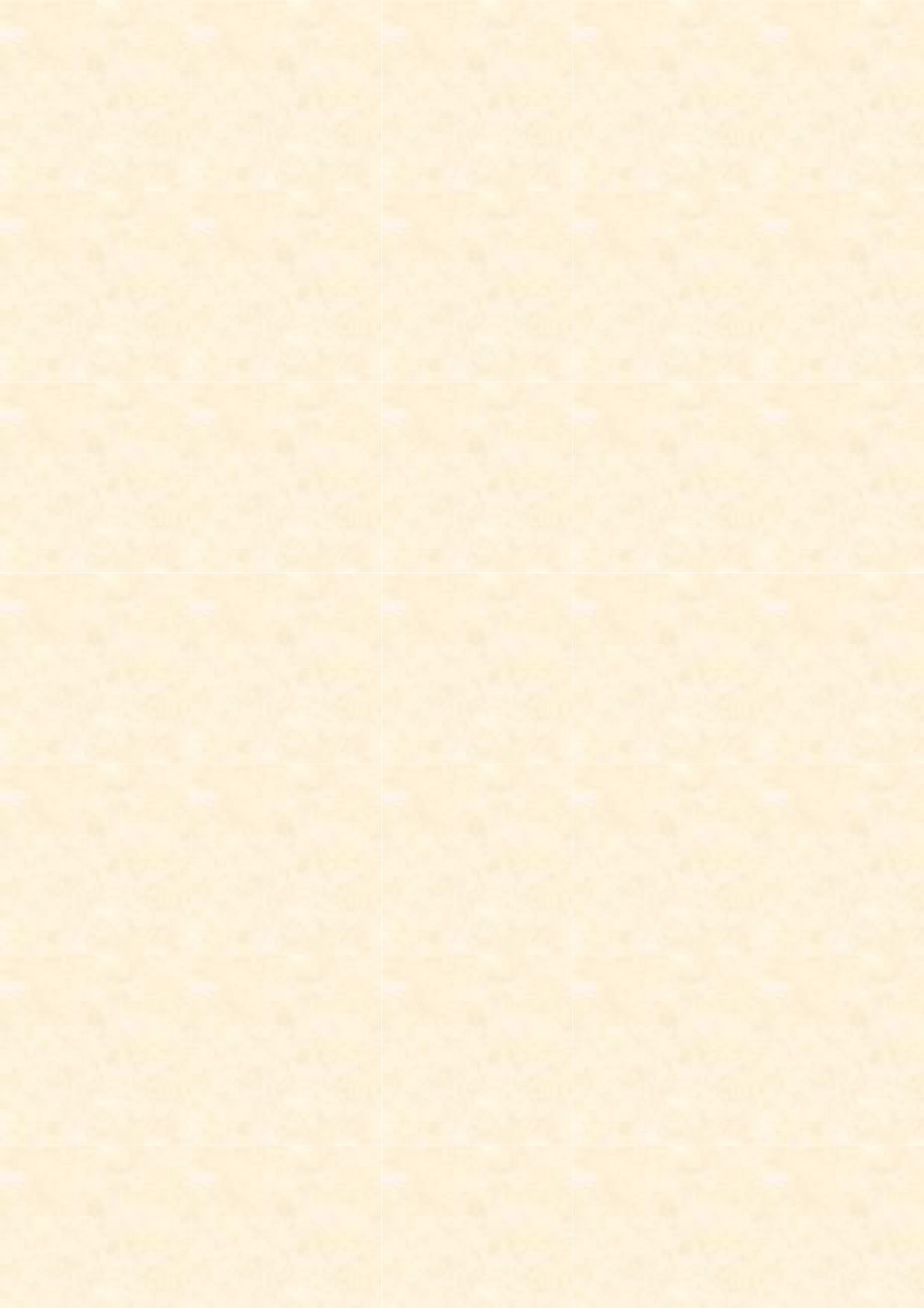
 Yah, Well !!!!

I think we need to Change as per Environment ….

As we’re Facing the Such Challenges ( Improving the system Configuration and Build New Infrastructure is not a good option , Because it’s a repetative Problem , It’ll last till end & increases rapidely) , we need to think new ….

1. **We can shift to Online Learning system , but we need to take Pain of Maintaining the Data Pipeline , to make sure that Good Input Data..**
2. **It’ll be Good Option to Build the pipeline Which avoid the Corrupted data , using**

our past learning knowledge , so that our Repetative problem will be solved ….

Q7: **How were you maintaining the failure cases?**

 **In this project, we have coded structure while keeping exception in mind. So, whenever there is any casuality and model doesn’t predict the correct output. Then the dataset get stored in the database. We have all the databricks connected to the azure pipeline. So, whenever there is any failure, we manually label the prediction and store back into database.**

Q8: **What kind of automation have you done for data processing**?

 **We have used full fledged ETL pipeline in this project. Banks already have data of credit card applicants. These all datasets are preprocessed through the steps followed while building the models, and later used for prediction using model.**

Q9 : Have you used any scheduler?

 **Yes, Job scheduler was used to run the databricks notebooks at a specific time.**

## Q10 : How are you monitoring your job?

**We have used Azure Pipeline for monitoring the run.**

### Q11. What were your roles and responsibilities in the project?

**My responsibilities involve preprocessing the dataset, applying feature selection, EDA, model training on the prepared dataset, deploying the trained model to the cloud, monitoring the deployed model for any issues, providing QA support before deployment and then providing the warranty support post-deployment.**

I was also working simultaneously with data engineers and data analysis team helping them to understand the data architecture and creating dashboards

Q12: **What was your day to day task?**

**My day to day tasks involved completing the JIRA tasks assigned to me, attending the scrum meetings, participating in design discussions and requirement gathering, doing the requirement analysis, data validation, image labeling, Unit test for the models, providing UAT support, etc.**

Q13 : What type of Preprocessing You have Done ?

 Handling the Missing Values , Feature scaling …

( all the features have Numerical data , Hence Part of Categorical data Handling is Absent)

Handling Missing Values : KNN imputer Feature Scaling : Standard Scaler

Q14 : **In which area you have contributed the most?**

 **I have worked mostly on EDA and model building part. As told earlier, I was also working simultaneously with other teams while working on the project.**

## Q15 :What are the Different types of Missing Value Imputation Techniques?

 1.Replacing it with mean , median , mode

 2.Knn Imputer

 **3.Imputation using most frequent( From sklearn.impute import SimpleImpute)**

## Q16 : What are the Different types of Outlier Detection Techniques ,?

1. **Box plot**
2. **Using Imperial Formulae**
3. **IQr techniques**
4. **Db scan Clustering**
5. **5 star methods…**
6. **Visualizing the Scatter plot**

**Q17 : What are the Different Task you perform in Feature Engineering**

**?**

 1.Handling the Missing Values

 2.Outlier detection

 3.Handling Categorical Data

 4.rescaling Operation ( If required)

 5.Variable Transformation

 Creating new Features.

## Q18 :Various Methods of the Feature Selection Techniques?

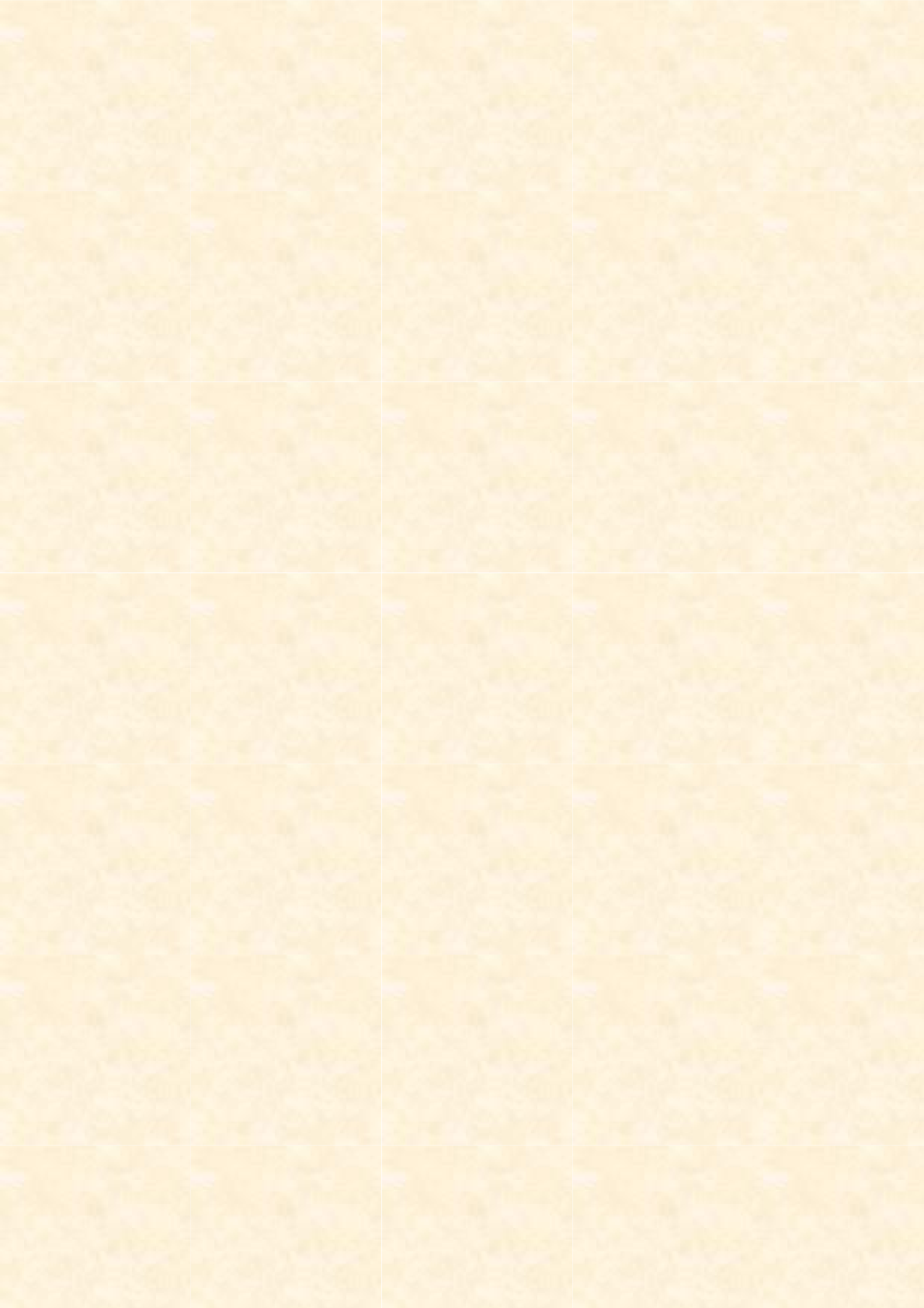
 1.Pearson Correlation

2.Chi2 test (k-best) 3.Backtracking method

4.Feature importance methods (tree based) 5.Annova

6.variance threshold methods 7.Fisher Score

1. **BackWord Elemination**
2. **info Gain (mutual\_info\_class) 10.Lasso Reg( )**

**Q19:** **In which technology you are most comfortable?**

 **I have worked in almost fields starting from Machine Learning, Deep Learning, NLP, Computer Vision, Chatbox and web development part.**

# Q20 : How you rate yourself in big data technology?

 **I have considerable knowledge in Big Data Technologies, but right now I am more focussed with working in Data Science. In big data I would like to rate myself 6, and in Data Science I would like to rate myself 7.5**

## Q21 : How were you doing deployment?

 **In this project we were using Azure APPservice for entire deployment of code through Git.**

# 

**Q22 : What kind of challenges have you faced during the project?**

**The biggest challenge we faced was during feature engineering journey of the datasets. There was a large number of features, and we have to perform various tests to take only useful measures for building the models.**

## Q23 : What are different Cross-Validation Techniques , Which one you used in your project & Why ?

1. K-fold , Leave One Leave out , Stratified k-fold , Grid search cv , Hold Out Grid Search Cv : To find out the Best Parameters for model ..

**k-fold :** *- helps to overcome the Overfitting , gives the clear idea about the model Performance*

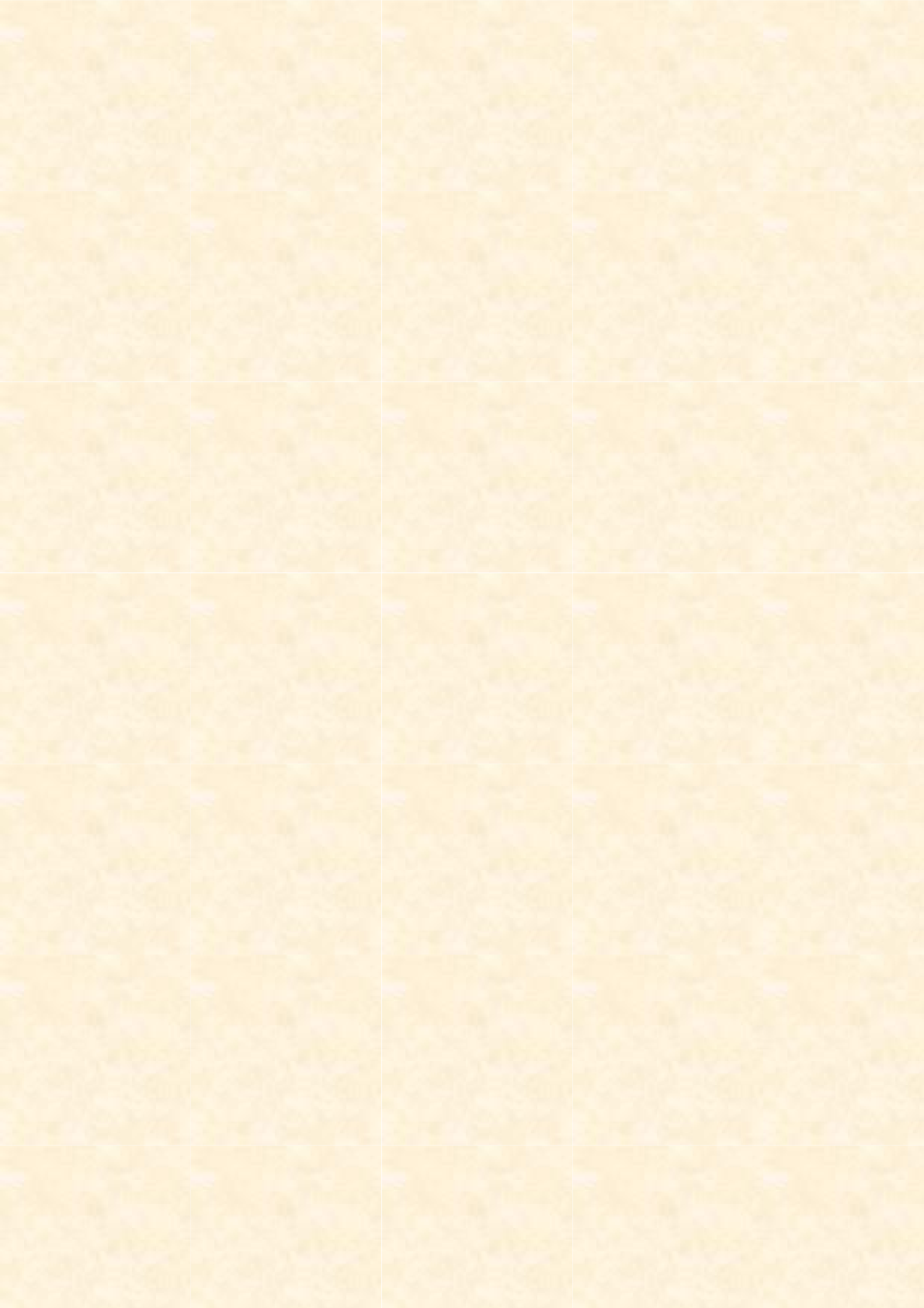
*- If the Range between highest score and lowest score is high , then it'll the sign of overfitting ,*

*we need todo hyper-prameter tunning again*

**stratified K-Fold :** *helps to overcome the Overfitting , gives the clear idea about the model Performance*

* *if dataset is unbalanced , then we use strati fied K-fold validation for the model validation ,*
* *StratifiedKFold , ensures that the equal propo rtion of element should be choosen from each class whil e cross-validation*
* *If the Range between highest score and lowest score is high , then it'll the sign of overfitting ,*

*we need todo hyper-prameter tunning again*

Hold Out Validation Approach :

***''' It's simple train test split dataset , divides datas et in two part(train & test )***

***- data selected randomly , as we change the Rand om\_state model Accuracy will change it's the***

***major drawback of this approch..***

***'''***

## In this project we have used , Grid Search cv , cross val score , hold out , stratified K-fold..

**Q24 : Which Performance Matrix You have Used in Project & Why , Talk more about the Various Performance Matrixes ?**

###  We’re reducing the FN .. ( trying to improve the sensitivity)

**No any faulty sensor will be detected as Normal …..**

**If we’re predicting (normal sensor as Faulty ) up to Some extent it’s Digestible , because the maintenance person can go at particular sensor Location & cross check it….**

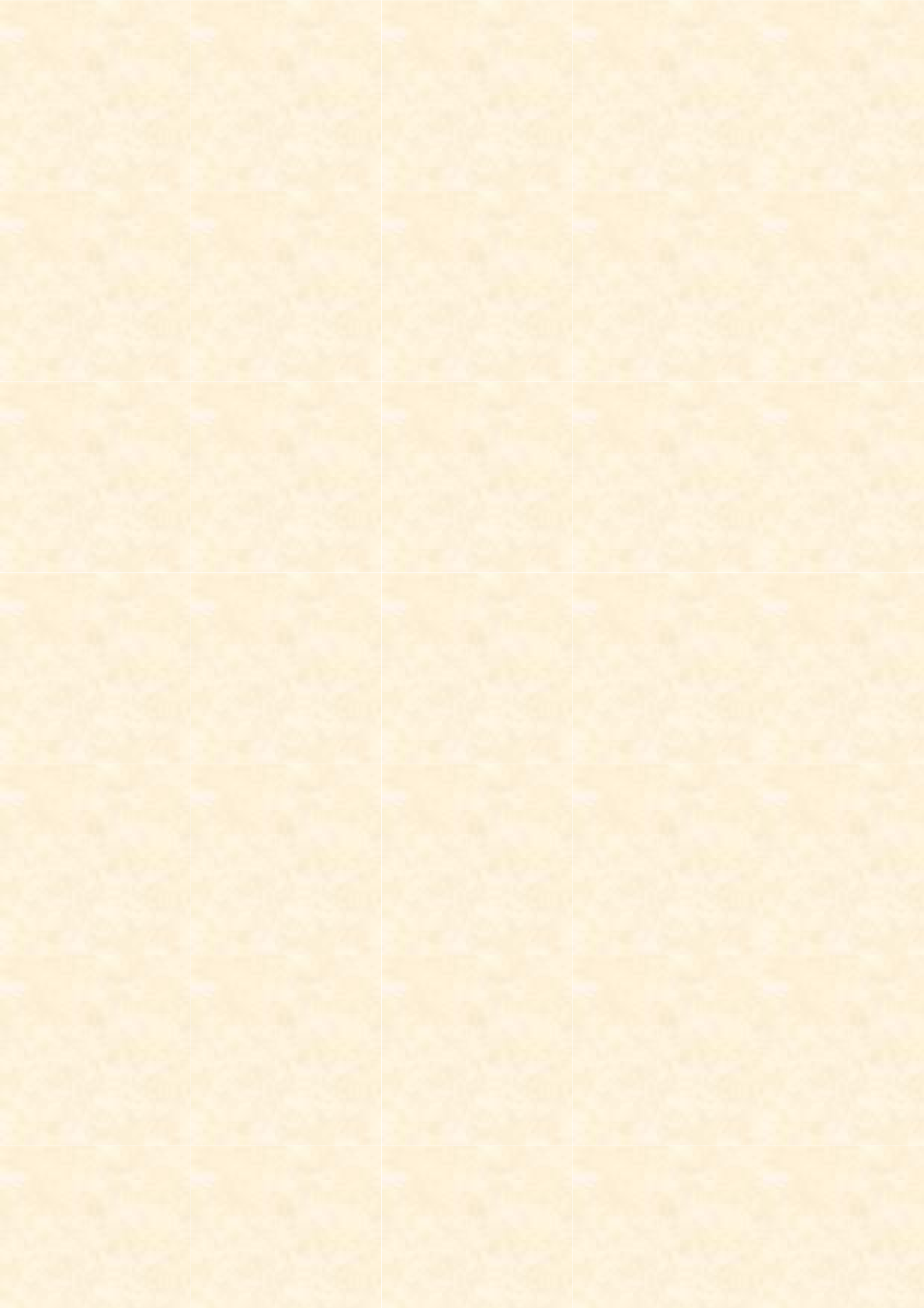
**And if we’re predicting the Faulty sensor as Normal Sensor ,**

**Then it’s again very difficult to track the Faulty sensor & it’s very time Consuming task…. And it’ll not be suitable for Our Project title…**

**Confusion\_matrix : talk abt Conf matrix ( FP , FN , TP , TN ) Sensitivity OR recall = TP / ( TP + FN ) ,**

**Precision = TP / ( TP + FP )**

**In this project we’re also using Accuracy , ROC and AUC curve**



## Q25: How you’re making Coding part more Generalized ,and make maintanence Easy ? ( talk about Logs)

 In this project , all the part of Coding Done in Modular fasion ..

Having some Advantages : Reusability , Easy to Understand , Easy to debug , Easy to make changes in particular Pipeline..

Maintaining the Logs : so that we can track the Internal Operations ,and we’ll be getting the current as well as historical info about run in text format ,, based on that we can Analyze & Optimized the Code & processes..

## Q25: What kind of Cloud Services You’re using( Cloud You’re Using ) ?

 **We’re using AZURE Cloud Services here…**

## Q26: How Frequently You decided to train model ?

1. To keep the model update we need to retrain the model on Old + new data So that system will be robust and more Generalized…

Basically , the retraining of model is depends on how frequently your data is changing ( rate of Change of data ) …

If rate of Change of( frequency ) is very high , then we need to retrain the model Rapidly ( once a day , week , once a month )

On other hand :

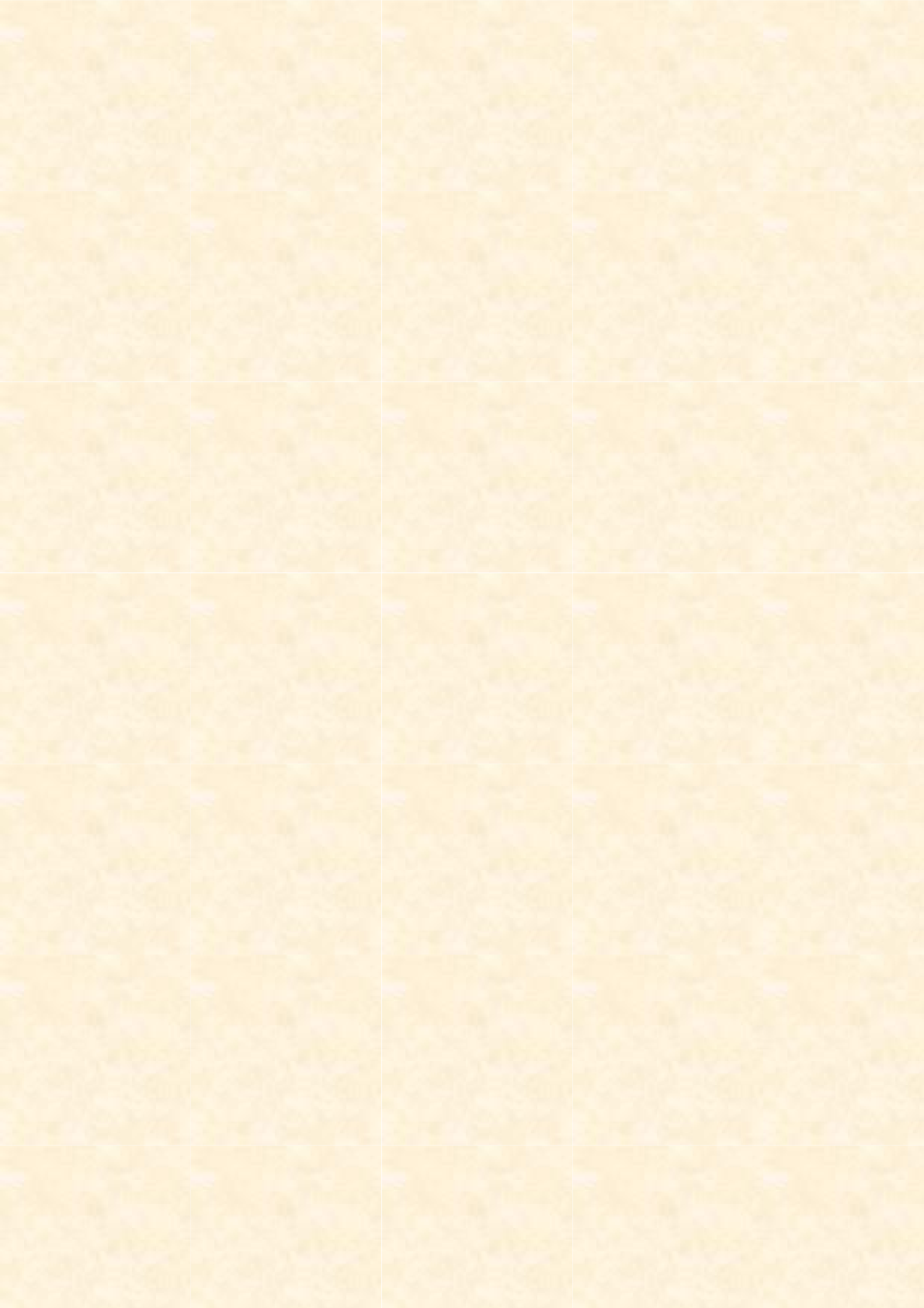
If rate of Change of( frequency ) is very less , then we not need to retrain the model Rapidly , we can retrain the model ( once a 2/3 months)

In our case , the Sensors data is not much depends on the Outside Factors , and hence very slightly changing day by day ..

And also if , industry is installing one more sensors ( or willing to change in electronic infrastructure ) then we need to modify our whole system also …

Hence , We can retrain the Our Models ( once 1 or 2 months)..

## 27 : What is the Team Size & Your role ?

 7 team member :

 Actually the role was not fixed , every week we get the task to perform and we

have to work on that task …

 Mostly I was active in installing data preprocessing pipeline , model pipeline building and deployment part ;

## Q28 : Problem involved in Project Duration ?

**Every week New task / Not fixed role : I learn to be a dynamic**

**from this activity ….**

**Aligning the team members . Maintaining the co-ordination in team , Handling such wide dataset , EDA ,**

**deployment part , tuning the model with such high dimensions..**

## Q29 : What kind of Clustering you have done and Why (k-mean)?

For this project I have used K means clustering technique which is a non hierarchial based clustering because we can define number of clusters we want to distribute the dataset for our model.

**Q30 : Which Algorithms used in Project , Why ( details about Algos) ?**

 **K-mean clustering**

 **Random Forest Classifier**

 **XGB Classifier**

**( Explain the Operation of Each )**

**Why these classifiers only ?**

1. **we tried with all algos but Random Forest and XGB performing better ..**

**Random Forest and XGB classifier can Handle the Wide data ….**

**Mention the advantages of each**