phase 5

<u>project documentation and submission</u>

name : S . Sheeba

naan mudhalvan ID : au920821104044

phase 5 : project submission and submission

customer churn prediction :

project objectives :

customer churn prediction means knowing which customers are likely to leave or unsubscribe from your service . for many companies , this is an important prediction . this is because acquiring new customers often costs more than retaining existing ones.

Design thinking process :

1.Empathize :

Understand the needs and pain points of the business : Gather insights into why customers are churning,such as poor customer service ,pricing issues, or product dissatisfaction .

2.Define :

Develop a concise problem statement that outlines the objective of the churn prediction system and the key metrices to be used.

3.ideate :

___ Brainstorm solutions: Encourage cross-functional teams to generate ideas for predicting churn.consider both traditional statistical models nd machine learning approaches .priortize ideas : Use criteria like feasibility ,potential impact,and cost effectiveness to rank and select the most promising ideas.

4.Test:

Implement the churn prediction system.Develop a full scale system based on the prototype ,integrating with relevant data source and existing business processes .continuously monitor the systems performance and refine itbased on real world data and feedback.

5.Implement:

Roll out the system:Deploy the churn prediction prediction systemto production and provide training to relevant personnel.Create action plans: Develop strategies and actions to be taken when the system predicts a customer is at risk of churning .

6.Evaluate:

Define key performance indicators(KPIs) to assess the system impact on reducing customer churn.Gather feedback: Collect feedback from customer service teams and customers to make iterative improvements .

7.Iterate:

__Continuously update and refine the churn prediction system based on the data and feedback collected .Adapt to changing customer behavior and market conditions to maintain the systems effectiveness.

Developement phases :

1.Data source:

Churn prediction relies on data from various sources,including senior citizen,gender, techsupport, phoneservice, multiple lines, internet service and

customer feedback

2. Data Preprocessing:

- Data preprocessing involves cleaning and transforming data to make it

suitable for analysis. This includes handling missing values, outliers, and

feature engineering.

3. Feature Selection:

- Identifying the most relevant features (customer attributes) is essential for

accurate churn prediction. Common features include customer lifetime value,

usage patterns, and customer support

interactions.

4. Model Building:

- Machine learning models, such as logistic regression, decision trees,

random forests, and neural networks, are used to build predictive models.

- Models are trained on historical data where the churn outcome is known.

5. Model Deployment:

- Once a reliable churn prediction model is developed, it can be integrated

into operational systems for real-time predictions.

- The model might trigger actions, such as sending retention offers or alerts to customer support teams.

Techniques :

1.Ensemble learning:

• Random forest model: Random forest models combine multiple

decision trees to reduce o verfitting and increase prediction accuracy

• Gradient boosting: Algorithms like XGboost,Light bgm,and catboost

use gradient boosting to build powerful predictive models

2. Feature engineering

• Create new features that capture customer behavior,such as customer

lifetime value,recency,frequency,and monetary value(RFM analysis)

3.Anamoly detection

• Identifying unusual customer behavior using techniques like isolation

forests or one class SVMS

4.Time-series analysis:

• Analyzing historical customer data as a time series to detect temporal

patterns in churn

5.Hyperparameter optimization

• Using techniques like Bayesian optimization or grid search to find the

best parameters for your models

6.Transfer learning:

• Leveraging pre-trained models on related tasks,such as

recommendation systems or customer

segmentations,to enhance

churn prediction

7.Model evaluation:

• Using advanced metrics like AUC-ROC,AUC-PR,or F1-score to assess

model performance,especially when dealing with imbalanced datasets

8.Imbalanced data handling:

• Tecniques like oversampling,undersampling,or synthetic data

geaneration to address class imbalance issues in churn prediction

9.Automl:

• Automated machine learning platforms can help automate the model

selection and hyperparameter tuning

process,making it easier to find

the best model for the specific churn prediction problem

10.Recurrent neural networks(RNNs):

• RNNs are used for sequence modeling,making them suitable for churn

prediction when dealing with time-series data

11.Data preprocessing:

• Data preprocessing involves cleaning and transforming data to make it

suitable for analysis and this includes handling missing

values,outliers,and feauture engineering

12.Feature selection:

Identifying the most relevant features is essential for accurate churn prediction and common features include customer lifetime value, usage

patterns, and customer support interaction

<u>Dataset link :</u>

https://www.kaggle.com/datasets/blastchar/telco-customer-churn

A	B	C	D	E	F	G	H	I	J	K	L	M	N
customerI	gender	SeniorCiti:	Partner	Dependen	tenure	PhoneSer	MultipleLi	InternetSe	OnlineSec	OnlineBac	DevicePro	TechSupp	Streaming
7590-VHV	Female	0	Yes	No	1	No	No phone	DSL	No	Yes	No	No	No
5575-GNV	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No
3668-QPY	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No
7795-CFO	Male	0	No	No	45	No	No phone	DSL	Yes	No	Yes	Yes	No
9237-HQI	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No
9305-CDSI	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes
1452-KIOV	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes
6713-OKO	Female	0	No	No	10	No	No phone	DSL	Yes	No	No	No	No

ABSTRACT:

• Customer churn, the rate at which

customers discontinue their association

with a company, poses a significant challenge for businesses across industries. In an era marked by data abundance, this study leverages the power of data analytics to predict and mitigate customer churn. This research employs a comprehensive dataset of customer interactions, including demographics, transaction history, and customer feedback, and applies various machine learning and statistical : techniques to develop predictive models. The aim is to identify the key factors that influence customer attrition and provide businesses with actionable insights to proactively retain their customer base. The results show promising predictive accuracy, offering companies an opportunity to optimize their customer retention strategies and enhance customer satisfaction. This research contributes to the growing field of customer relationship management by showcasing the potential of data analytics in predicting and preventing customer churn, ultimately fostering sustainable business growth.

DATA SOUCE:

• Churn prediction relies on data from various sources,including senior citizen,gender, techsupport, phoneservice, multiple lines, internet service and customer feedback

<u>code:</u>

```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.ticker as mtick
import matplotlib.pyplot as plt
data="C:/churn/Telco-Customer-Churn.csv"
df=pd.read_csv(data)
print(df)


print(df.head())
print(df.info())


#step3:data preprocessing
df.columns.values
df.dtypes
df.isnull().sum()

# removing missings values

df.dropna(inplace=True)

#removing customer IDs from the dataset
df2=df.iloc[:,1:]

#converting the predictor variable to a binary numeric variable
df2.replace(to_replace='yes',value=1,inplace=True)
df2.replace(to_replace='no',value=0,inplace=True)

#let's convert all the categorial variables into dummy variables
```

```python
df.isnull().sum()

# removing missings values

df.dropna(inplace=True)

#removing customer IDs from the dataset
df2=df.iloc[:,1:]

#converting the predictor variable to a binary numeric variable
df2.replace(to_replace='yes',value=1,inplace=True)
df2.replace(to_replace='no',value=0,inplace=True)

#let's convert all the categorial variables into dummy variables

df_dummies=pd.get_dummies(df2)
df_dummies.head()

#get correlation of churn with other variables
plt.figure(figsize=(15,8))
df_dummies.corr().sort_values(ascending=False).plot(kind='bar')

# data explortion
colors=['#4D3425','#E4512B']
ax=(df['gender'].value_counts()*100.0/len(df)).plot(kind='bar',stacked=True,
ax.yaxis.set_major_formatter(mtick.percentFormatter())
ax.set_ylabel('%customers')
ax.set_xlabel('gender')
ax.set_ylabel('%customer')
ax.set_title('Gender Distribution')
```

```python
# Senior citizen

ax = (df['SeniorCitizen'].value_counts() * 100.0 / len(df)).plot.pie(autopct='%.1f%%',
                                                      labels=['No', 'Yes'],figsize=(5, 5), fontsize=12)


ax.yaxis.set_major_formatter(mtick.PercentFormatter())
ax.set_ylabel('Senior Citizens', fontsize=12)
ax.set_title('% of Senior Citizens', fontsize=12)
```

```python
# Partner and dependent status

df2 = pd.melt(df, id_vars=['customerID'], value_vars=['Dependents', 'Partner'])
df3 = df2.groupby(['variable', 'value']).count().unstack()
df3 = df3 * 100 / len(df)
colors = ['#4D3425', '#E4512B']
ax = df3.loc[:, 'customerID'].plot.bar(stacked=True, color=colors, figsize=(8, 6), rot=0, width=0.2)

ax.yaxis.set_major_formatter(mtick.PercentFormatter())
ax.set_ylabel('% Customers', size=14)
ax.set_xlabel('')
ax.set_title('% Customers with dependents and partners', size=14)
ax.legend(loc='center', prop={'size': 14})

for p in ax.patches:
    width, height = p.get_width(), p.get_height()
    x, y = p.get_xy()
    ax.annotate('{:.0f}%'.format(height), (x + 0.25 * width, y + 0.4 * height),
                color='white',
                weight='bold',
                size=14)
```

```python
# Customers with or without dependents

colors = ['#4D3425', '#E4512B']
partner_dependents = df.groupby(['Partner', 'Dependents']).size().unstack()

ax = (partner_dependents.T * 100.0 / partner_dependents.T.sum()).T.plot(kind='bar',
                                                                         width=0.2,
                                                                         stacked=True,
                                                                         rot=0,
                                                                         figsize=(8, 6),
                                                                         color=colors)


ax.yaxis.set_major_formatter(mtick.PercentFormatter())

ax.legend(loc='center', prop={'size': 14}, title='Dependents', fontsize=14)
ax.set_ylabel('% Customers', size=14)
ax.set_title('% Customers with/without dependents based on whether they have a partner', size=14)
ax.xaxis.label.set_size(14)
```

```python
# Churn by Monthly Charges

ax = sns.kdeplot(df.MonthlyCharges[(df["Churn"] == 0)],
                 color="Red", shade=True)
ax = sns.kdeplot(df.MonthlyCharges[(df["Churn"] == 1)],
                 ax=ax, color="Blue", shade=True)
ax.legend(["Not Churn", "Churn"], loc='upper right')
ax.set_ylabel('Density')
ax.set_xlabel('Monthly Charges')
ax.set_title('Distribution of monthly charges by Churn')
```

```python
# Churn by Total Charges

ax = sns.kdeplot(df.TotalCharges[(df["Churn"] == 0)],
                 color="Red", shade=True)
ax = sns.kdeplot(df.TotalCharges[(df["Churn"] == 1)],
                 ax=ax, color="Blue", shade=True)
ax.legend(["Not Churn", "Churn"], loc='upper right')
ax.set_ylabel('Density')
ax.set_xlabel('Total Charges')
ax.set_title('Distribution of total charges by Churn')
```

```python
ax = sns.distplot(df[df['Contract'] == 'One year']['tenure'],
                  hist=True, kde=False,
                  bins=int(180 / 5), color='steelblue',
                  hist_kws={'edgecolor': 'black'},
                  kde_kws={'linewidth': 4},
                  ax=ax2)
ax.set_xlabel('Tenure (months)', size=14)
ax.set_title('One Year Contract', size=14)

ax = sns.distplot(df[df['Contract'] == 'Two year']['tenure'],
                  hist=True, kde=False,
                  bins=int(180 / 5), color='darkblue',
                  hist_kws={'edgecolor': 'black'},
                  kde_kws={'linewidth': 4},
                  ax=ax3)

ax.set_xlabel('Tenure (months)')
ax.set_title('Two Year Contract')

services = ['PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity',
            'OnlineBackup', 'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies']

fig, axes = plt.subplots(nrows=3, ncols=3, figsize=(15, 12))
for i, item in enumerate(services):
    if i < 3:
        ax = df[item].value_counts().plot(kind='bar', ax=axes[i, 0], rot=0)
```

```python

fig, axes = plt.subplots(nrows=3, ncols=3, figsize=(15, 12))
for i, item in enumerate(services):
    if i < 3:
        ax = df[item].value_counts().plot(kind='bar', ax=axes[i, 0], rot=0)

    elif i >= 3 and i < 6:
        ax = df[item].value_counts().plot(kind='bar', ax=axes[i - 3, 1], rot=0)

    elif i < 9:
        ax = df[item].value_counts().plot(kind='bar', ax=axes[i - 6, 2], rot=0)
    ax.set_title(item)

df[['MonthlyCharges', 'TotalCharges']].plot.scatter(x='MonthlyCharges', y='TotalCharges')
```

conclusion :

Future work include incorporating these future work considerations will help maintain the effectiveness and relevance of your customer churn prediction system ,ensuring its continued contribution to the success of your business.