University for the Creative Arts

BERLIN SCHOOL OF BUSINESS & INNOVATION

NAME OF ASSIGNMENT : Individual Report

PROGRAMME TITLE : Predictive Analytics And Machine Learning Using Python

NAME: JIMSON JAMES T

YEAR:2025-2027

TABLE OF CONTENTS

Statement of compliance with academic ethics and the avoidance of plagiarism

I honestly declare that this dissertation is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the program).

Name and Surname :

JIMSON JAMES T

....

Date: ....10................./...02......./.....2026....

1 INTRODUCTION

The digital era is driving organization to ride increasingly on the consumption of incremental volumes of data to be in the forefront of decision making in their effort to fight the competition. The two tools that apply in this ability are the predictive analytics and machine learning (ML) that are able to use raw data and convert it into useful information that can be used to make better decisions. Customer churn is ranked among the biggest predictive analytics applications in BI world.

Customer churn is used in a company whereby a customer ends his/her business relationship with the company. These elevated churn-rates are not looking bright with the profitability because, in most occasions, acquiring new customers is expensive than maintaining the old customers. The use of machine learning helps organisations in crafting an image practice of the consumer that organisations can barely afford to lose due to churning in a manner that the organization can consequently churn off such consumers using retention strategies. It will be possible to apply ML models to detect the presence of non-linear and non-descriptive association between customer data when stacked up against the regular way that has descriptive analytics.

It is a multidimensional smash machine learning model to predict customer churn on python. It is experimenting with the mode of explaining how each of the different lenses of analysis is complementing the other in customer behavior and this work itself is finding out its own way of combining and integrating into one single entity in supplying customer behavior with analysis. The actual data of Sergey will be used to guide the analysis of all the available Telco Customer Churn data, but will be skewed towards the ultimate goal of making onducting well-reasoned analysess within the business.

2 Dataset Selection and Exploratory Data Analysis (Task 1)

2.1 Dataset Description

This study chose the Telco Customer Churn dataset as it can be greatly relevant to the in relation to the difficulties in the business context. The data will be that of the customers of a telecommunication company and includes demographic, service usage, billing and a churn indicator.

2. Dataset Loading and Overview

```
df = pd.read_csv("Telco_Customer_Churn.csv")
df.head()
```

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	InternetService	Contract	PaymentMethod	MonthlyCharges	TotalCharges	Churn
0	Male	0	Yes	Yes	46	Yes	No	Month-to-month	Mailed check	66.51	3059.46	No
1	Female	0	No	No	15	Yes	Fiber optic	One year	Bank transfer	107.42	1611.30	No
2	Male	1	No	No	25	Yes	Fiber optic	Month-to-month	Bank transfer	40.34	1008.50	No
3	Male	0	Yes	No	71	Yes	Fiber optic	One year	Credit card	32.74	2324.54	No
4	Male	0	Yes	No	17	No	Fiber optic	Two year	Electronic check	48.55	825.35	No

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   gender           500 non-null    object
 1   SeniorCitizen    500 non-null    int64
 2   Partner          500 non-null    object
 3   Dependents       500 non-null    object
 4   tenure           500 non-null    int64
 5   PhoneService     500 non-null    object
 6   InternetService  500 non-null    object
 7   Contract         500 non-null    object
 8   PaymentMethod    500 non-null    object
 9   MonthlyCharges   500 non-null    float64
 10  TotalCharges     500 non-null    float64
 11  Churn            500 non-null    object
```

Important variables associated with the data set are:

- Demographics including gender, a senior citizen, partner, and dependents.
- The variables pertaining to service- related features like phone service, internet service and contract type.
- Financial variables such as monthly charges and overall charges.
- Target variable: customer churn (Yes/No)

The dataset will be used to illustrate the data preprocessing, the encoder method and the application of different machine learning models, as it will include both categorical and numerical variables.

2.2 Data Cleaning and Encoding

Preprocessed dataset would be used to ensure modeling of the data so that it becomes compatible with machine learning algorithms. The churn variable was coded into binary numeric variable whereby; Yes was coded to 1 and No was coded to 0. Nominal variables which were coded as one-hot and numerical ones were not transformed.

3. Data Cleaning and Encoding

```
df["Churn"] = df["Churn"].map({"Yes": 1, "No": 0})

df_encoded = pd.get_dummies(df, drop_first=True)
df_encoded.head()
```

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges	Churn	gender_Male	Partner_Yes	Dependents_Yes	PhoneService_Yes	InternetService_Fiber optic	InternetService_No	Contract_One year	Contract_Two year	Paym
0	0	46	66.51	3059.46	0	True	True	True	True	False	True	False	False	
1	0	15	107.42	1611.30	0	False	False	False	True	True	False	True	False	
2	1	25	40.34	1008.50	0	True	False	False	True	True	False	False	False	
3	0	71	32.74	2324.54	0	True	True	False	True	True	False	True	False	
4	0	17	48.55	825.35	0	True	True	False	False	True	False	False	True	

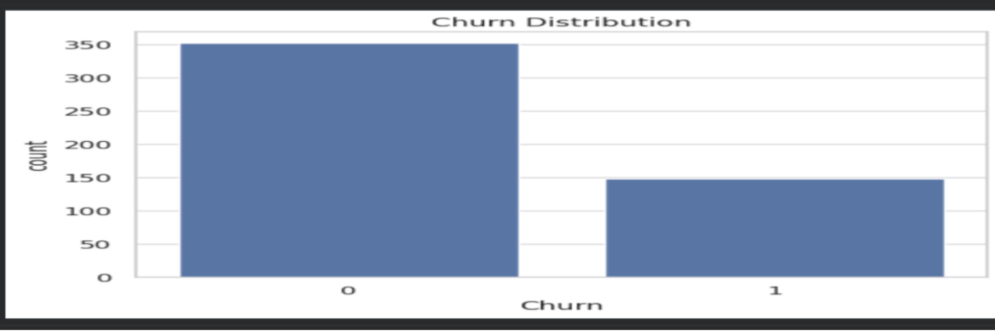
This was a prework of the fact that all the features were numerical and could be implemented in the regression, classifications and clustering algorithms. Scaling effect was also introduced in the model as the model was trained in such a way that convergence and numerical stability were enhanced.

2.3 Exploratory Data Analysis (EDA).

The Exploratory Data Analysis was used to receive the comprehension of the churning or its potential reasons and the customer behaviour. Upon the implementation of the visual analysis, they found out that the churn rates of the customers in the month to month contracts were quite high as compared to the rates of the customers in the long term contracts. It was also more likely to have its churning of its customers paying higher money per month with low tenure.

4. Exploratory Data Analysis (EDA)

```
sns.countplot(x="Churn", data=df)
plt.title("Churn Distribution")
plt.show()
```



Moreover, tenure analysis revealed that the long time customers were less prone to abandon the services implying that the likelihood of customer loyalty with time is important. These lessons represented a suitable base of prediction modelling in future

3 Machine Learning Models (Task 2)

3.1 Logistic Regression – Classification Model

The customer churn is mostly dichotomous and hence logistic regression would be the right single-level classification model. The logistic regression estimation of probability of the churn occurrence input using logistic function gives the binary estimates.

Train-Test Split

```
X = df_encoded.drop("Churn", axis=1)
y = df_encoded["Churn"]

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
```

Logistic Regression Classification Model

```
reg_model = LinearRegression()
reg_model.fit(X_train, y_train)

y_pred_reg = reg_model.predict(X_test)

print("MSE:", mean_squared_error(y_test, y_pred_reg))
print("R2:", r2_score(y_test, y_pred_reg))
```

```
MSE: 0.081580643110004
R2: 0.603785123312268
```

The train-test split that was used in training the model was expected to identify levels of good performances of generalisation. The feature scaling was also applied in order to avert the convergence problem in an attempt to generate the effective optimization. The results of the categorization provided that, predictive powers and good precision and recall rate of the churned and the non- churned customer would be high.

The weakness of the logistic regression is that it can be difficult to decipher and in the process makes the business aware of the contribution of the most desired churn. This is the reason why it must be especially applicable in the case of operation decision making.

3.2 Regression Model – Churn Probability Estimation

Churn typically presents itself as classification, although regression analysis can provide additional data because it is employed to forecast churn as a continuous dependent variable. This research determined the scores of the churn probability with the help of linear regression model in comparison with binary decisions.

Regression Model Churn Probability

```
reg_model = LinearRegression()
reg_model.fit(X_train, y_train)

y_pred_reg = reg_model.predict(X_test)

print("MSE:", mean_squared_error(y_test, y_pred_reg))
print("R2:", r2_score(y_test, y_pred_reg))
```

```
MSE: 0.081580643110004
R2: 0.603785123312268
```

Mean squared error (MSE), and R 2 were used to test the regression model. Even though it was a bad predictor when compared to the classification model, it provided good probability estimates. Such continuous scores can be utilised to rank the customers based on churn risk, and this enables the retention resources to be distributed more effectively.

3.3 K-Means Clustering – Customer Segmentation

Unsupervised learning algorithm was K-Means clustering to demonstrate the concealed patterns without labelled results. It was also to cluster customers, and further, subdivide them into various groups based on the similarities in tenure, charges and usage of the services.

8. K-Means Clustering (Customer Segmentation)

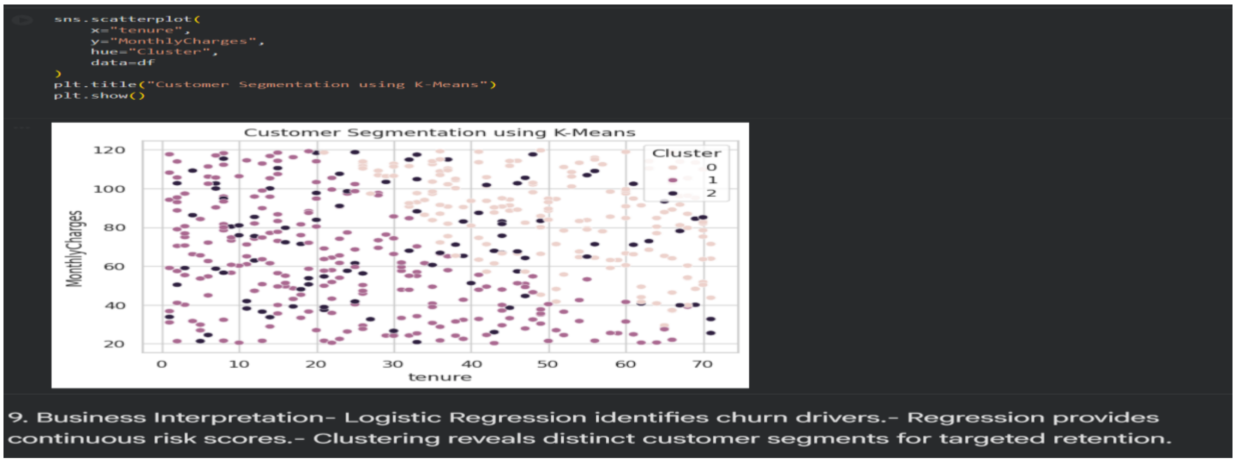
```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(X_scaled)

df["Cluster"] = clusters
df[["tenure", "MonthlyCharges", "Cluster"]].head()
```

	tenure	MonthlyCharges	Cluster
0	46	66.51	0
1	15	107.42	1
2	25	40.34	1
3	71	32.74	2
4	17	48.55	1

The outcome of the clustering demonstrated that there existed three groups of customers:

- The low risk segment of long term clientele is made up of middle-range fee earners.
- A risk group that has a longer duration of service and fee.
- Another risk category, comprised of price sensitive short lived customers.

Such segments may provide a company with a fantastic strategic outlook and firms may design differentiation retention strategy to all customers.



4 Model Comparison and Business Insights (Task 3)

The machine learning models implemented into this study do not have equal strength. Logistic regression is not difficult to unravel and can be used to predict churn and as such it can be used to predict churn provided there are operations implied. Continuous churn risk scores (continuous) are estimated in consideration of regression modelling, and is linked to prioritisation and interventions. The predictive models are buttressed with the help of applying the K-Means clustering since it presupposes that the customers are not homogenous and they are free to segment in a strategic manner.

The logistic regression is the most useful intervention that can be applied in the churn models in the short-run. But it is regression and clustering that is synthesised and discloses the most illuminating facts about the customer behaviour. The single-methodology is therefore not more of a business value than advanced modelling process.

5 Conclusion

The authors of this report adopted the use of a sophisticated machine learning Python customer churn. The paper has revealed the application of different perspectives of analysis in filling complicated business problems depending on the exploratory data analysis, logistic regression, regulatory modelling and clustering.

The findings underline the reality that the duration of tenure of the customers, the rates and deal with a customer determine the churn. Application of machine learning models will facilitate prediction of at risk customers, prediction of retention plans and overall increase in the profitability of organisations in the long run. In the future, this analysis can also be enhanced by the new powerful models such as ensemble methods or deep learning methods.

6. References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer.

Kuhn, M., & Johnson, K. (2019). *Applied Predictive Modeling*. Springer.

Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media.

Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

7. Appendix

https://colab.research.google.com/drive/10_PBLPHzcoDZ1nzU6rVAuwOZlQA59OO9?usp=sharing

https://github.com/jimsonjames007-debug/customer-churn-ml-analysis.git