Data Science Project

by

Opitien Brume Pascal

R Studio and Power BI

INTRODUCTION

This project focuses on understanding customer behaviour in the telecommunications industry using data from a fictional company named Bangor Telco. The main challenge for Bangor Telco and many other companies in this industry is keeping their customers, which is also known as reducing customer churn.

To tackle this problem, I will first employ three popular data analysis methods: Decision Trees, Logistic Regression, and k-nearest Neighbors (kNN). Each of these methods will help predict which customers might leave the company (churn) and understand different groups of customers based on their behaviour and characteristics.

Furthermore, an essential part of this project is creating a Data Science Dashboard. This tool will use the model to show how well the predictions work and allow users to input information and get predictions. This dashboard is not just for showing the results; it is a practical tool for the company to make data-driven decisions. By tackling this project, I aim to provide valuable insights to Bangor Telco.

These insights can help them improve their services, make their customers happier, and reduce the number of customers leaving the company. This project shows how data science can be used in real-world situations to help businesses understand their customers better and make smarter decisions.

DATA COLLECTION

In this section, I connected to Bangor Telco's MySQL database to retrieve customer data crucial for the analysis by using the RMySQL package in R to establish a connection using essential credentials: username, password, host, database name, and port number.

To retrieve the data, the SQL query "SELECT * FROM customer_churn.customers" is executed to retrieve all records from the customer table. After data retrieval, the database connection is closed to maintain security and resource efficiency.

```
   CUSTOMERID  COLLEGE  INCOME  OVERAGE  LEFTOVER   HOUSE  HANDSET_PRICE
1 BTLC-007761     zero   89318        0         0  162233            266
2 BTLC-007682      one  142814      187        17  346690            716
3 BTLC-002228     zero   55675        0        32  792662            257
4 BTLC-011752      one   39559        0         0  416439            165
5 BTLC-015958     zero  145081        0         0  341108            583
  OVER_15MINS_CALLS_PER_MONTH  AVERAGE_CALL_DURATION  REPORTED_SATISFACTION
1                           1                     12                  unsat
2                          24                      4                  unsat
3                           1                      1            very_unsat
4                           0                     15              very_sat
5                           0                      9                    avg

  REPORTED_USAGE_LEVEL         CONSIDERING_CHANGE_OF_PLAN          LEAVE
1                very_little                        considering    STAY
2                       high                         considering  LEAVE
3                very_little                      never_thought    STAY
4                       high                        considering    STAY
5                        avg                                  no  LEAVE
```

DATA PREPROCESSING

In this step, I refined the dataset to ensure its suitability for analysis. The process involved:

1. Renaming Columns: For better readability and ease of analysis column names are standardized to reflect the data they represent.

2. Data Transformation: Variables are transformed into categorical variables (factors) and the required binary format for certain analytical models.

3. Data Integrity Checks: Checks for missing values (nulls) and found no such issues in the data set.

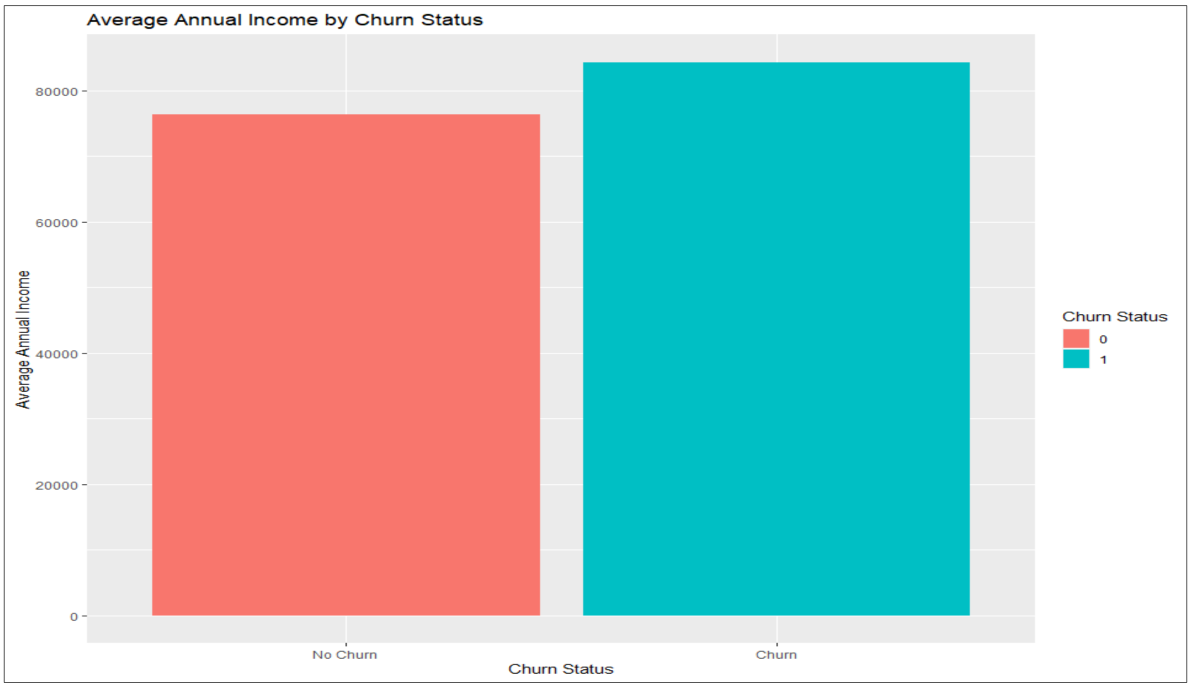
```
               customer_id                    college              annual_income
                         0                          0                          0
        monthly_overcharge  leftover_minutes_percent                house_value
                         0                          0                          0
                phone_cost        long_calls_per_month            avg_call_duration
                         0                          0                          0
        satisfaction_level                usage_level  considering_plan_change
                         0                          0                          0
              churn_status
                         0
```

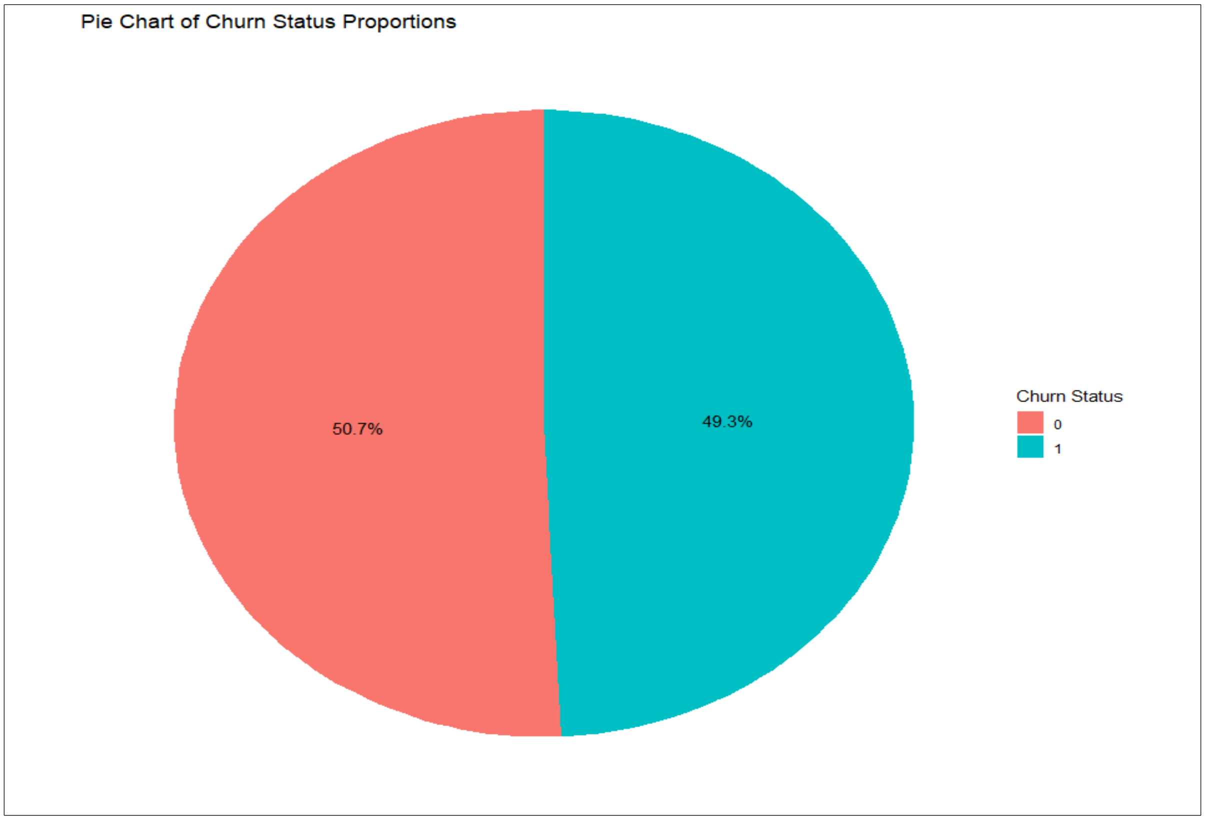
The variables have been successfully renamed and no nulls or duplicates are found.

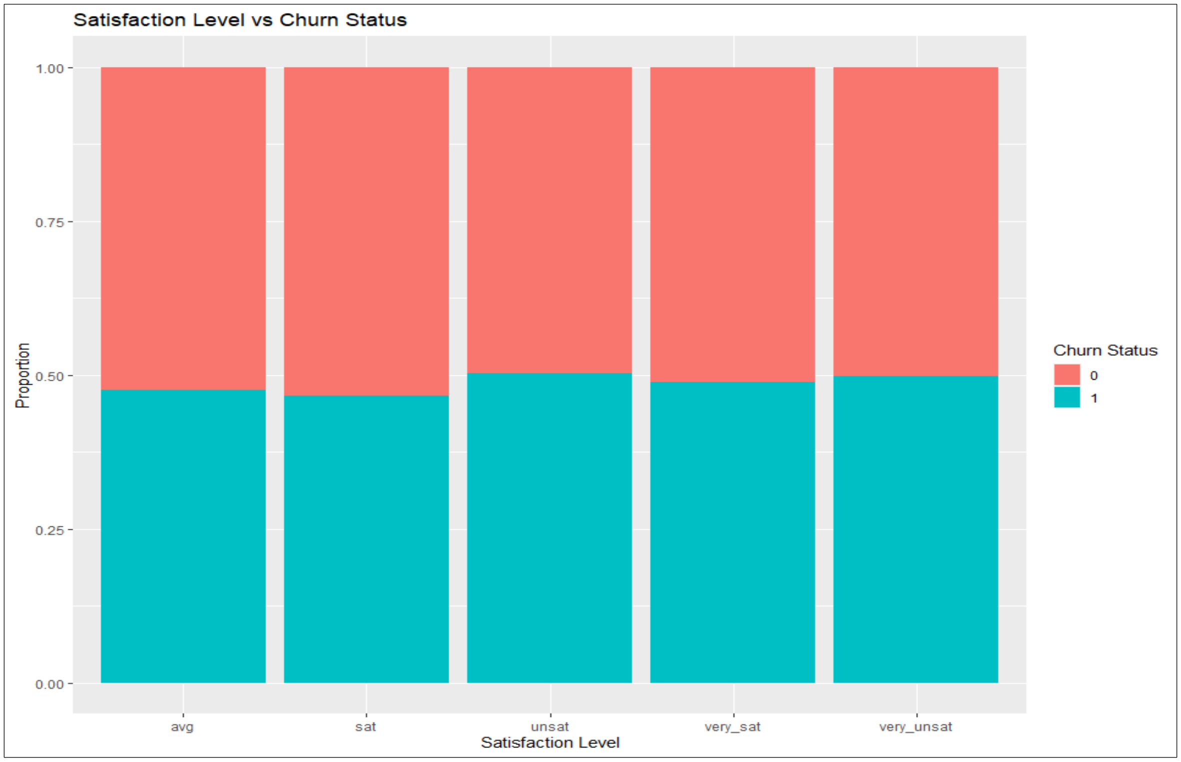
EXPLORATORY DATA ANALYSIS

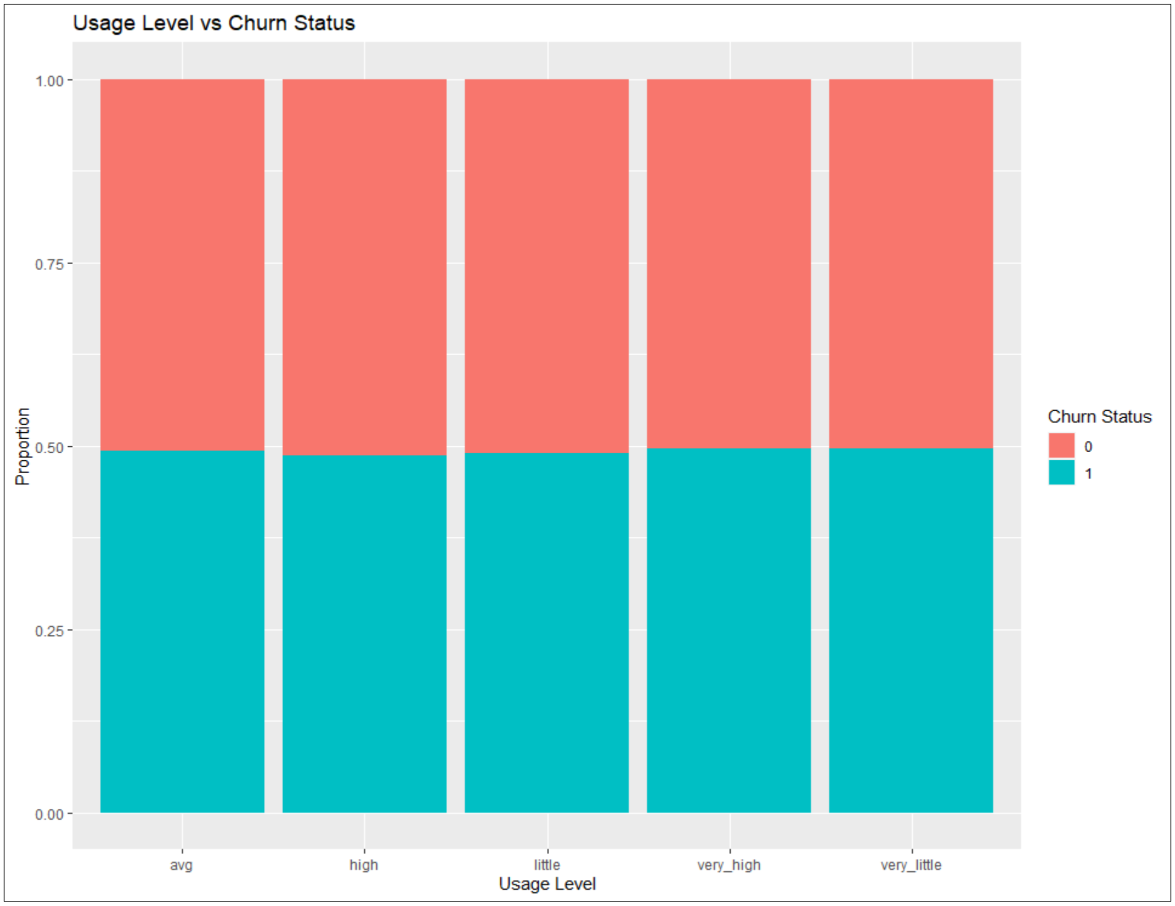
This section is focused on the Exploratory Data Analysis (EDA), which is important to understand the data characteristics and prepare for a more detailed analysis (A. S. Rao, 2021).

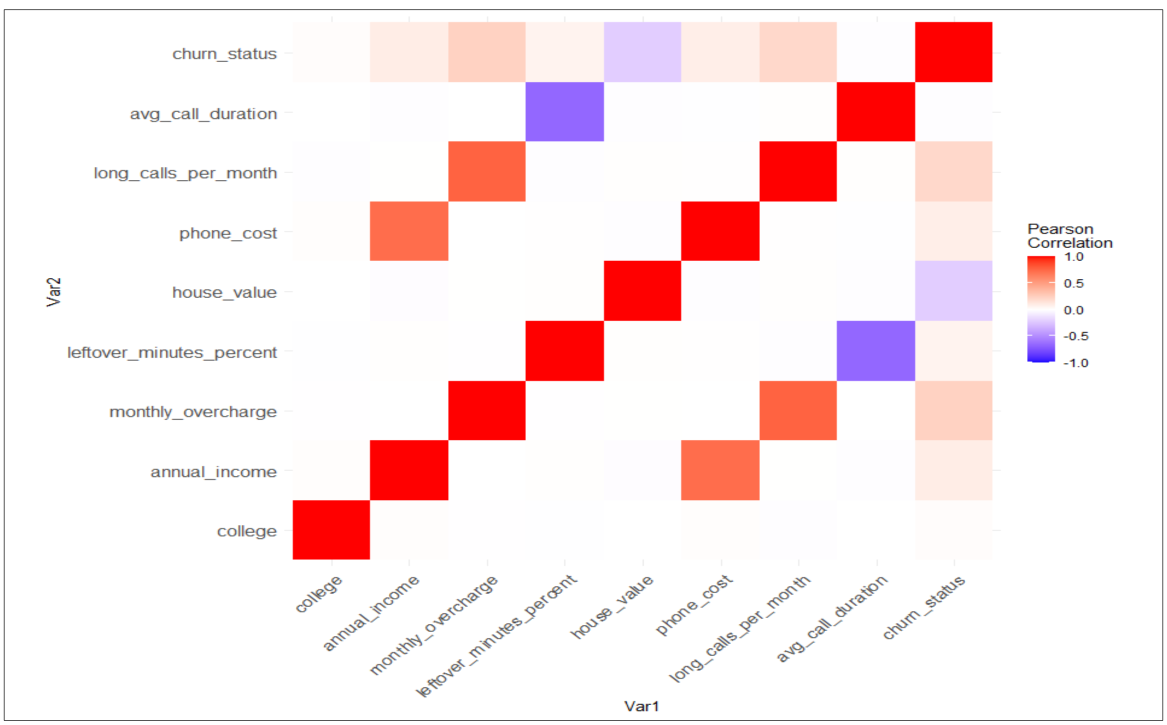
```
[1] "Churn Rate: 49.26 %"
```




Pie Chart of Churn Status Proportions
50.7%
49.3%
Churn Status
0
1


Satisfaction Level vs Churn Status
1.00
0.75
0.50
0.25
0.00
Proportion
avg
sat
unsat
very_sat
very_unsat
Satisfaction Level
Churn Status
0
1


Usage Level vs Churn Status
1.00
0.75
0.50
0.25
0.00
Proportion
avg
high
little
very_high
very_little
Usage Level
Churn Status
0
1


churn_status
avg_call_duration
long_calls_per_month
phone_cost
house_value
leftover_minutes_percent
monthly_overcharge
annual_income
college
Var2
college
annual_income
monthly_overcharge
leftover_minutes_percent
house_value
phone_cost
long_calls_per_month
avg_call_duration
churn_status
Var1
Pearson
Correlation
1.0
0.5
0.0
-0.5
-1.0

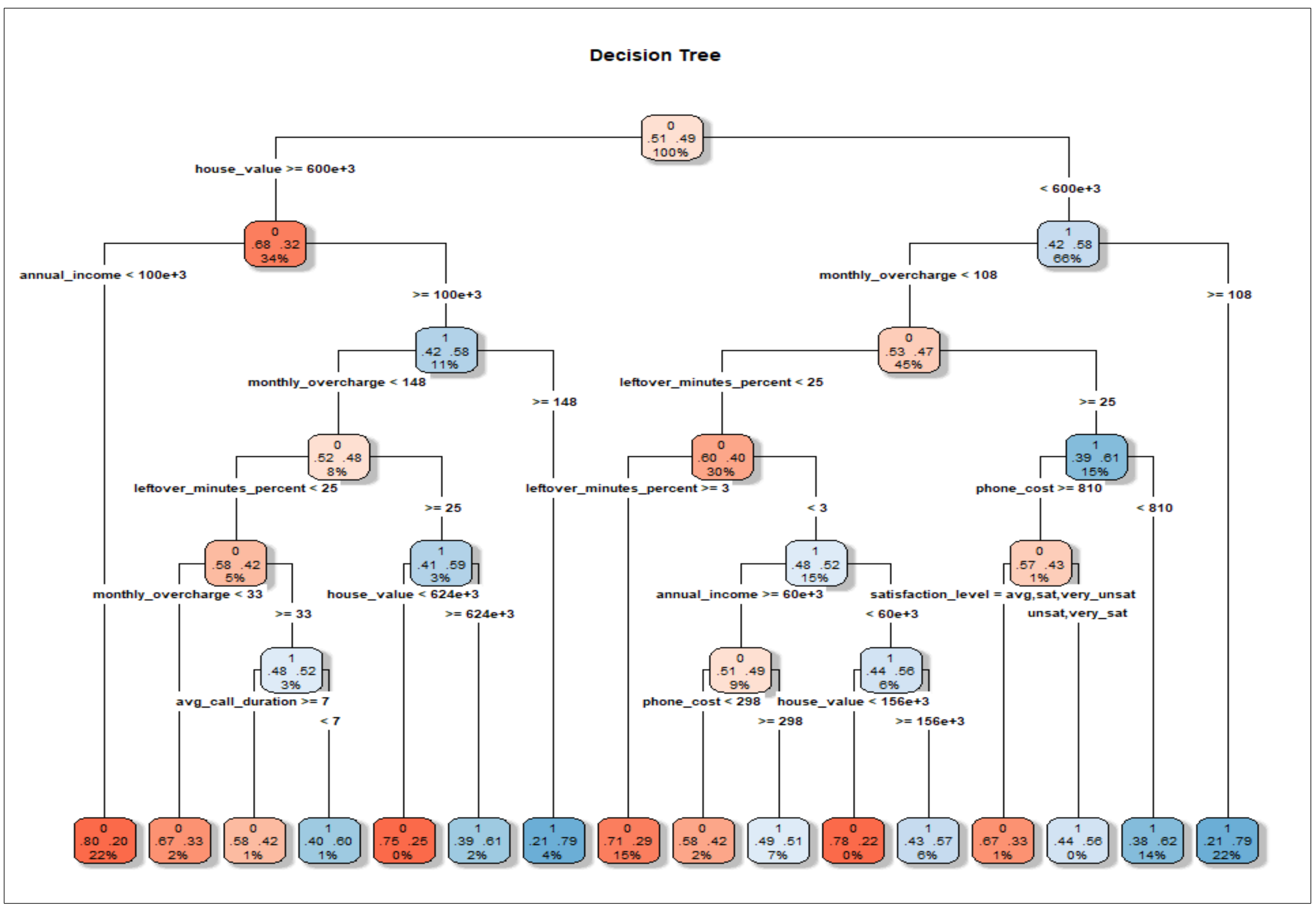
From the charts above, the data has a fair balance between the two categories; churn(1) and stayed(0). This is important to prevent bias in the models. Also, there is no issue with multicollinear variables from the correlation heatmap visual. Having delved into the data to identify insights, I will proceed to the model-building tasks.

MODEL BUILDING ANALYSIS

TASK 1: DECISION TREE CLASSIFICATION

The goal of this task is to develop a Decision Tree model which is a straightforward method to predict customer churn and understand the factors influencing their decisions. Decision Trees are one of the most popular approaches for representing classifiers. It is used in machine learning and data analysis that represents a procedure for computing the outcome of a function (Blockeel, et al., 2023). It involves performing tests on input data, with each test's outcome determining the next step until the function's result is known with certainty (Blockeel, et al., 2023).

```
Confusion Matrix:

predictions    0    1
          0 1989  651
          1 1055 2305

Model Evaluation Metrics:

Accuracy: 0.7157

Precision: 0.6860

Recall: 0.7798

F1 Score: 0.7299
```



Metrics

The Decision Tree model exhibits a strong predictive performance with an accuracy of 71.57%. This suggests that the model correctly predicts customer churn status in roughly 72 out of every 100 cases. The precision score of around 68.60% implies that when the model predicts a customer will churn, it is correct approximately 69% of the time. The recall score of approximately 77.98% indicates that the model successfully identifies about 78% of actual churn cases. Finally, the F1 score which balances between precision and recall is 72.99%, this shows the model’s robustness in accounting for both false positives and false negatives. These metrics collectively affirm the model's efficacy in predicting the churn likelihood among customers.

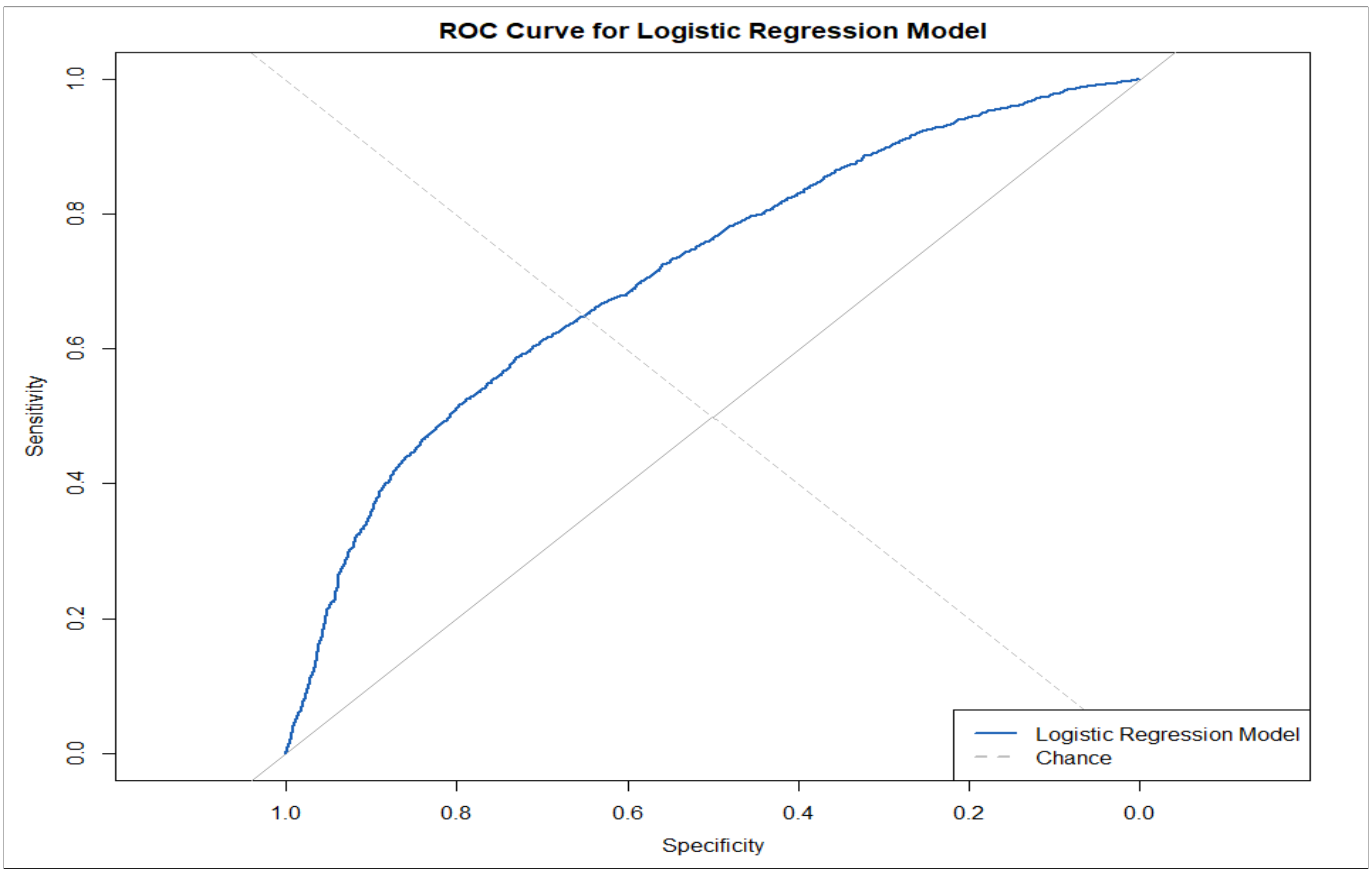
Tree Visual

The decision tree visual provided represents the model that classifies customer churn based on the key attributes. It initiates at the root node with a house value threshold and progresses into branches, symbolizing different decision paths. Each internal node represents a decision point, such as the annual income or monthly overcharge, and then splits the data further leading to the leaf nodes. The leaves represent the outcomes of customer churn indicating the count of customers predicted to churn (1) or not (0).

TASK 2: LOGISTIC REGRESSION

Logistic regression is a statistical method primarily used for binary and multiclass classification problems. It is designed to discover the relationship between a dependent variable and one or more independent variables and predict a binary outcome based on those variables (Stoltzfus, 2011).

This task aims to build the best logistic regression model to predict the probability a given customer will leave. To enhance the predictive performance of the model, I utilized the glmnet package. This package is distinguished for its regularization techniques which mitigate overfitting—a common issue when numerous predictors are present (Friedman, et al., 2010). Glmnet effectively reduces the complexity of the model while ensuring that it captures genuine patterns rather than noise. The cross-validation feature within glmnet further helps in selecting the best fit to generalize new and unseen data (Geeks, 2023).



```
Model Evaluation Metrics:

Accuracy: 0.6553

Precision: 0.6611

Recall: 0.6164

F1 Score: 0.6380

Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

```
Area Under the ROC Curve (AUC): 0.7114


Coefficients at Minimum Lambda:

22 x 1 sparse Matrix of class "dgCMatrix"
                                               s1
(Intercept)                         -3.968338e-01
college                              4.606091e-02
annual_income                        3.253468e-06
monthly_overcharge                   4.967890e-03
leftover_minutes_percent             7.300779e-03
house_value                         -1.735252e-06
phone_cost                           3.504767e-04
long_calls_per_month                 9.967076e-03
avg_call_duration                    2.109209e-02
satisfaction_levelavg               -2.828754e-02
satisfaction_levelsat               -1.420209e-01
satisfaction_levelunsat              4.119267e-03
satisfaction_levelvery_sat           .
satisfaction_levelvery_unsat         .
usage_levelhigh                      .
usage_levellittle                   -3.796469e-03
usage_levelvery_high                 8.622152e-04
usage_levelvery_little               5.470125e-04
considering_plan_changeconsidering   .
considering_plan_changenever_thought .
considering_plan_changeno            .
considering_plan_changeperhaps       .
```

The logistic regression model gives an accuracy of 0.6553, a fair capability to distinguish between customers who will stay and those likely to churn. The precision measure at 0.6611 suggests that the model has a modest rate of correctly predicting churn when it does so. The recall rate of 0.6164 shows that the model also has moderate success in identifying actual instances of churn among all potential cases. Finally, the f1 Score of 0.638 confirms that the model strikes a balance between precision and recall, yet it implies there is potential for improvement even after cross-validation.

The ROC curve has an area under the curve (AUC) of 0.7114 showing that the model has a good ability to differentiate between the two classes outperforming random chance. However, the curve and the scores suggest that further refinement could give more accurate predictions or perhaps an alternative algorithm.

Coefficients:

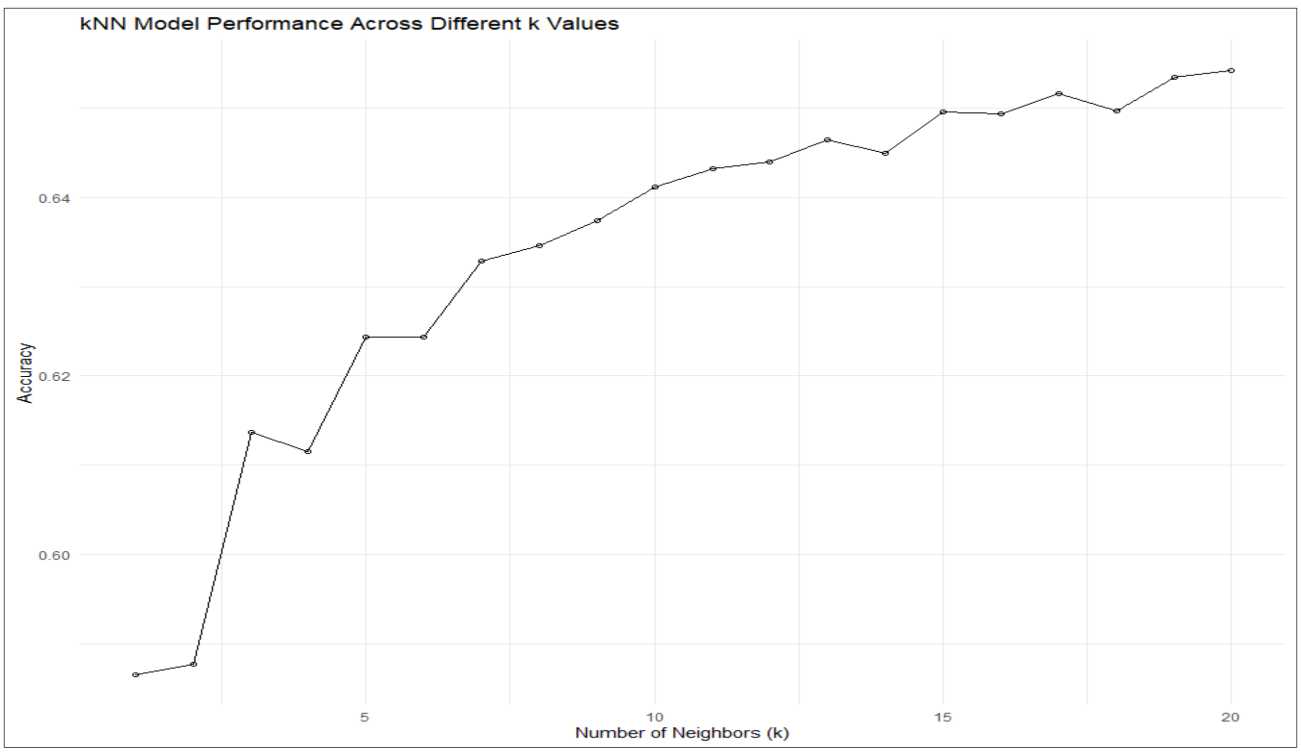
The coefficients are in a sparse matrix format. Each row represents a feature and the number in the s1 column is the coefficient for that feature at the selected lambda value (lambda. min). Coefficients close to zero have been effectively removed from the model by the regularization process. Furthermore, features with a dot (.) as the coefficient have been shrunk to zero by the regularization process, indicating they were not found to be significant predictors in the model by regularization applied. Finally, the features with non-zero coefficients are those that the model finds most statistically significant for predicting the outcome.

TASK 3: K NEAREST NEIGHBOR MODEL

The k-Nearest Neighbors (kNN) algorithm is a machine learning method used for both classification and regression tasks. It operates by assigning an object to a class most common among its k nearest neighbours in the training dataset (Zhongheng, 2016)

In Task 3 I addressed the challenge of predicting customer churn using the k-Nearest Neighbors (kNN) algorithm. It involved using the customer data to construct a predictive model that identifies the likelihood of customers discontinuing their service. I have carried out thorough data preparation, normalization, and the application of kNN to derive meaningful insights that inform retention strategies. The kNN model serves as a tool to capture patterns within potential churn.

```
Initial kNN Model Metrics:
Accuracy: 0.6507
Precision: 0.6508
Recall: 0.6279
F1 Score: 0.6391
```



Final kNN Model Metrics with k = 20 :

Accuracy: 0.6723

Precision: 0.6809

Recall: 0.6302

F1 Score: 0.6546

The aim of this task is to predict customer churn with KNN. The first step is normalizing numerical attributes to ensure consistent data scaling crucial for the distance-based kNN algorithm. The initial kNN model showed an accuracy of 65.07%, a precision of 65.08%, a recall of 62.79%, and an F1 score of 63.91%. It showed room for improvement therefore, to improve the model's robustness, a 10-fold cross-validation was applied which determined the optimal number of neighbours to be 20.

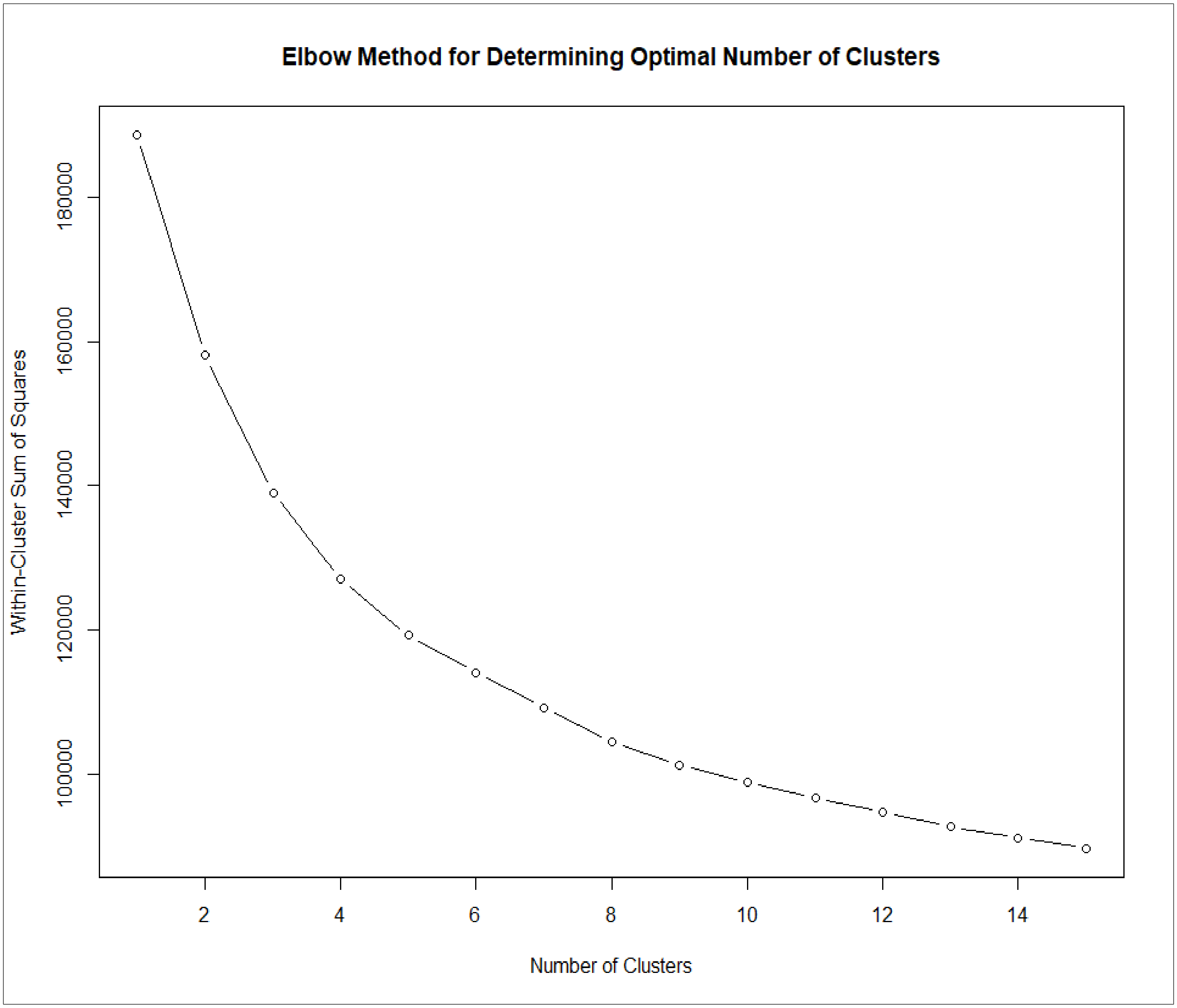
This adjustment enhanced accuracy to 67.23%, with a corresponding increase in precision to 68.09% and F1 score to 65.46%. The benefits of cross-validation are evident in the final model with improved metrics; consistency and reliability.

TASK 4 : CLUSTERING WITH K-MEANS

The k-means algorithm is a widely used method for clustering data. It involves partitioning a dataset into k distinct non-overlapping clusters and each point in the dataset is assigned to the cluster with the nearest mean, serving as a prototype of the cluster (Youguo & Haiyan, 2012).

This task aims to discover natural groupings within the customer dataset using clustering algorithms to identify distinct segments within the customer base which are not defined by pre-existing labels. These clusters are instrumental in informing targeted marketing strategies like offering tailoring service offerings. My goal is to leverage these insights to drive business decisions focused on enhancing customer satisfaction and retention.

```
  college annual_income monthly_overcharge leftover_minutes_percent house_value
1       0     0.2168048         -0.9998514               -0.8911872  -1.3110615
2       1     1.5002801          1.1747612               -0.2572525  -0.5802721
3       0    -0.5903575         -0.9998514                0.3021016   1.1865982
4       1    -0.9770124         -0.9998514               -0.8911872  -0.3039377
5       0     1.5546699         -0.9998514               -0.8911872  -0.6023871
6       1     0.9680659         -0.2323411               -0.2572525  -0.1004100
  phone_cost long_calls_per_month avg_call_duration satisfaction_level.avg
1 -0.5781251           -0.7843554      1.3623330904                      0
2  1.5264421            1.7925547     -0.4547924522                      0
3 -0.6202164           -0.7843554     -1.1362145307                      0
4 -1.0504835           -0.8963950      2.0437551688                      0
5  0.9044256           -0.8963950      0.6809110119                      1
6  2.3121472           -0.4482367     -0.0005110666                      0
  satisfaction_level.sat satisfaction_level.unsat satisfaction_level.very_sat
1                      0                        1                           0
2                      0                        1                           0
3                      0                        0                           0
4                      0                        0                           1
5                      0                        0                           0
6                      1                        0                           0
  satisfaction_level.very_unsat usage_level.avg usage_level.high
1                             0               0                0
2                             0               0                1
3                             1               0                0
4                             0               0                1
5                             0               1                0
6                             0               0                0
  usage_level.little usage_level.very_high usage_level.very_little
1                  0                     0                       1
2                  0                     0                       0
3                  0                     0                       1
4                  0                     0                       0
5                  0                     0                       0
6                  0                     1                       0
  considering_plan_change.actively_looking_into_it
1                                                0
2                                                0
3                                                0
4                                                0
5                                                0
6                                                0
  considering_plan_change.considering considering_plan_change.never_thought
1                                   1                                     0
2                                   1                                     0
3                                   0                                     1
4                                   1                                     0
5                                   0                                     0
6                                   1                                     0
  considering_plan_change.no considering_plan_change.perhaps
```



Elbow Method for Determining Optimal Number of Clusters
Within-Cluster Sum of Squares
100000
120000
140000
160000
180000
2
4
6
8
10
12
14
Number of Clusters

	cluster	college	annual_income	monthly_overcharge	leftover_minutes_percent
1	1	0.4898457	-0.3575727	-0.4719263	1.364899573
2	2	0.5079078	1.1377559	1.2460693	-0.003230461
3	3	0.5147508	1.1810469	-0.6231647	-0.186162031
4	4	0.4925555	-0.5822155	1.2716786	-0.143980038
5	5	0.5066777	-0.5698066	-0.6223431	-0.611962044

	house_value	phone_cost	long_calls_per_month	avg_call_duration
1	0.033669309	-0.4159486	-0.4996602	-0.95600272
2	-0.004407468	1.1830163	1.2458239	0.01160037
3	-0.015184757	1.2516744	-0.6143015	0.06089548
4	-0.002286780	-0.5983940	1.2701227	0.05781259
5	-0.007503979	-0.5856358	-0.6100614	0.49934826

	satisfaction_level.avg	satisfaction_level.sat	satisfaction_level.unsat
1	0.10018955	0.05009477	0.2079610
2	0.09850881	0.05558066	0.2019883
3	0.10325534	0.05289929	0.1904883
4	0.09470344	0.04759580	0.2003905
5	0.10552350	0.05177246	0.1988458

	satisfaction_level.very_sat	satisfaction_level.very_unsat	usage_level.avg
1	0.2548064	0.3869483	0.04874086
2	0.2435608	0.4003615	0.05286941
3	0.2606816	0.3926755	0.05518820
4	0.2457896	0.4115206	0.04808396
5	0.2540808	0.3897774	0.04682605

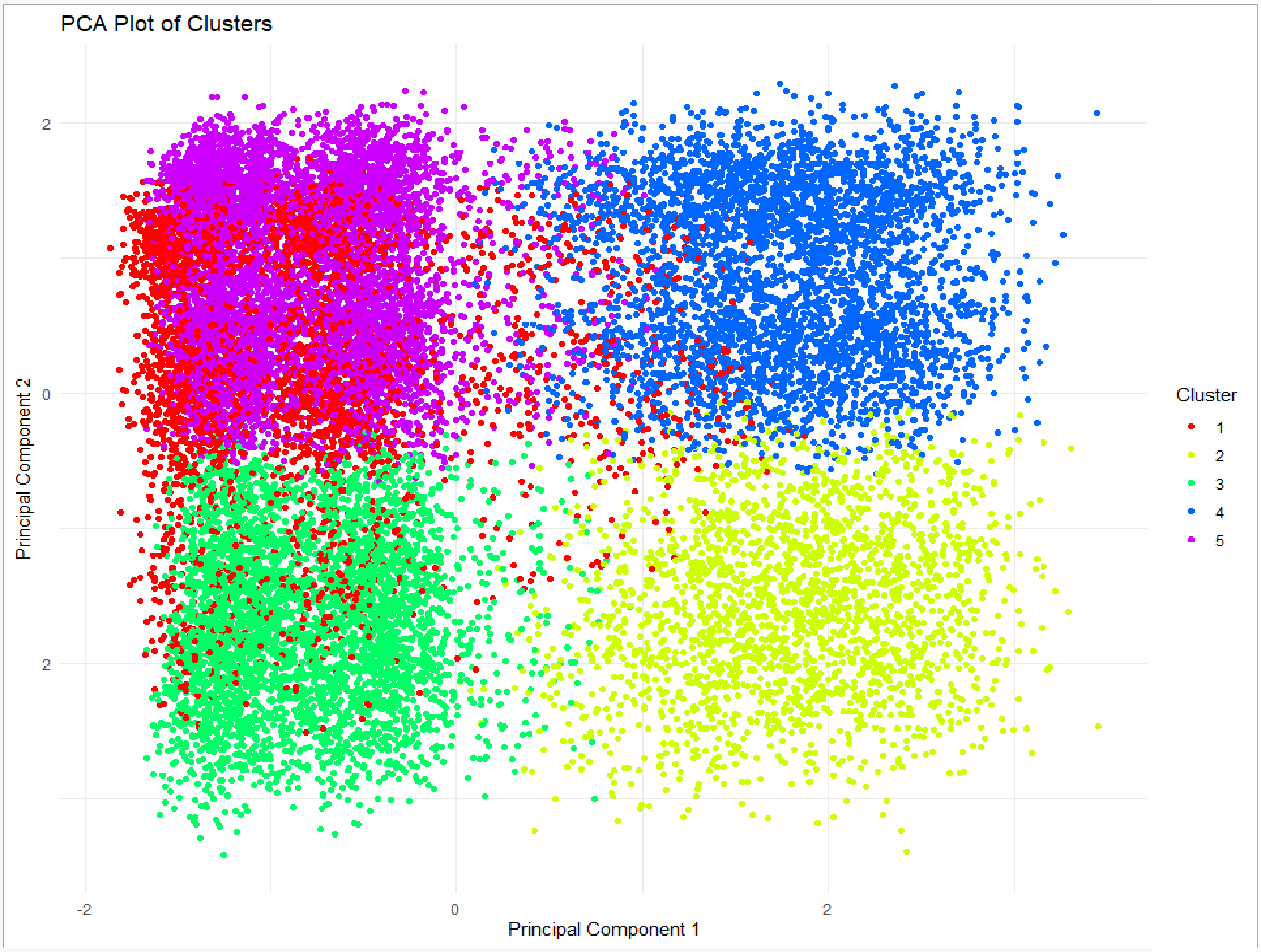
	usage_level.high	usage_level.little	usage_level.very_high
1	0.09694016	0.3964257	0.2539940
2	0.09986444	0.3994577	0.2489833
3	0.09689725	0.3916582	0.2555951
4	0.10226995	0.3941909	0.2528680
5	0.10239077	0.3910965	0.2603462

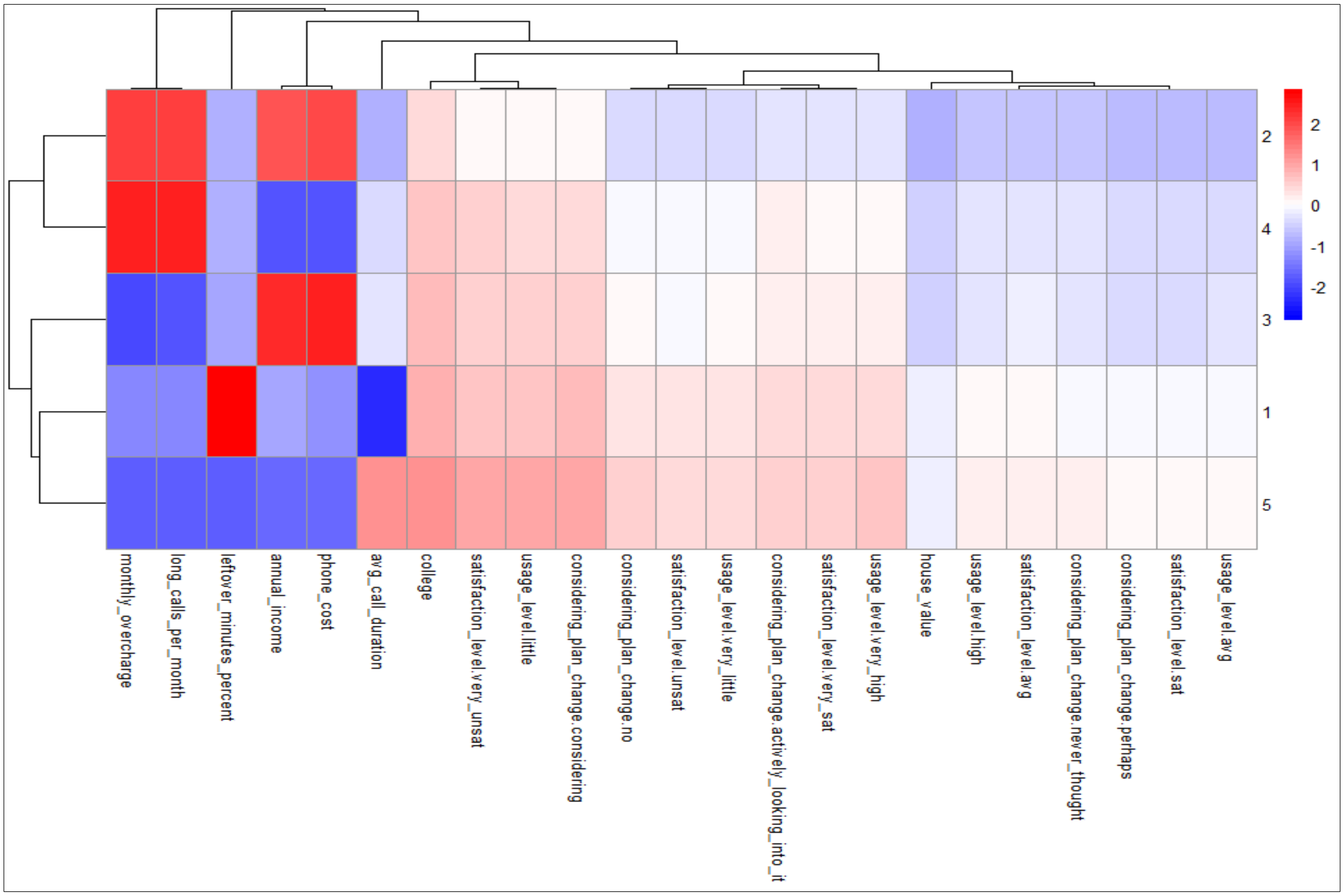
	usage_level.very_little	considering_plan_change.actively_looking_into_it
1	0.2038993	0.2499323
2	0.1988251	0.2498870
3	0.2006612	0.2525432
4	0.2025873	0.2640957
5	0.1993405	0.2379225

	considering_plan_change.considering	considering_plan_change.never_thought
1	0.4053615	0.09125372
2	0.4008134	0.09941256
3	0.3919125	0.10020346
4	0.3971198	0.09567977
5	0.3904369	0.10750206

	considering_plan_change.no	considering_plan_change.perhaps
1	0.1990252	0.05442729
2	0.1997289	0.05015816
3	0.2011699	0.05417091
4	0.1872102	0.05589456
5	0.2148392	0.04929926

1	2	3	4	5
3693	2213	3932	4097	6065





The k-means clustering results are illustrated through the PCA plot and the heat map. The PCA plot displays the distribution of data points across the principal components categorized by their clusters. The table above also shows the size of each cluster and provides a quantitative perspective on the segments within the customer base.

Cluster 2 Characteristics

- High annual income
- High monthly overcharge
- Moderate leftover minutes per cent
- Slightly below-average house value
- Very high phone cost
- High number of long calls per month
- Slightly above-average call duration
- Satisfaction levels are relatively distributed with a slight lean towards higher satisfaction.

- Moderate usage levels with a tendency towards higher usage
- Moderate consideration of plan changes

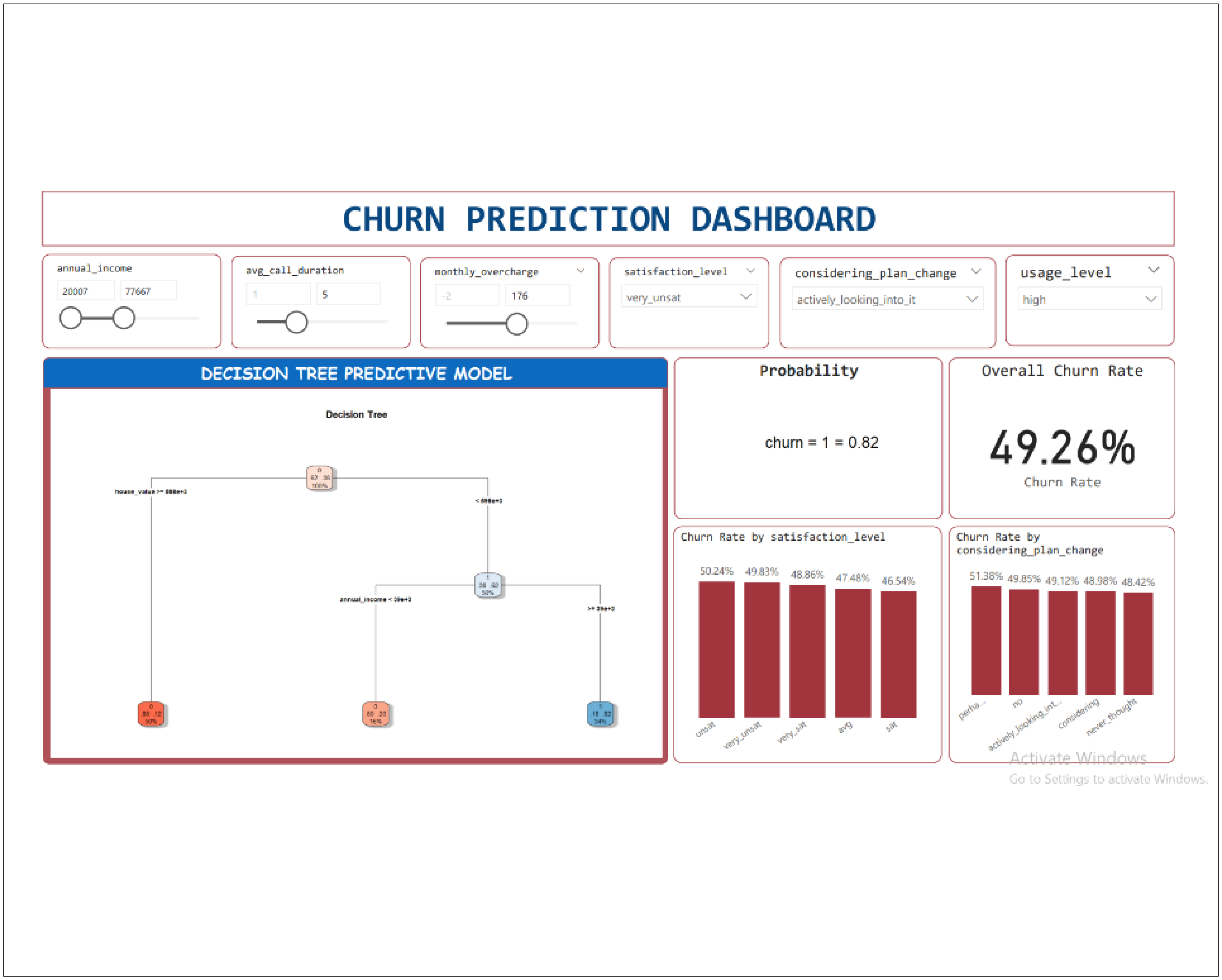
Explanation in Business Terms based on the Task:

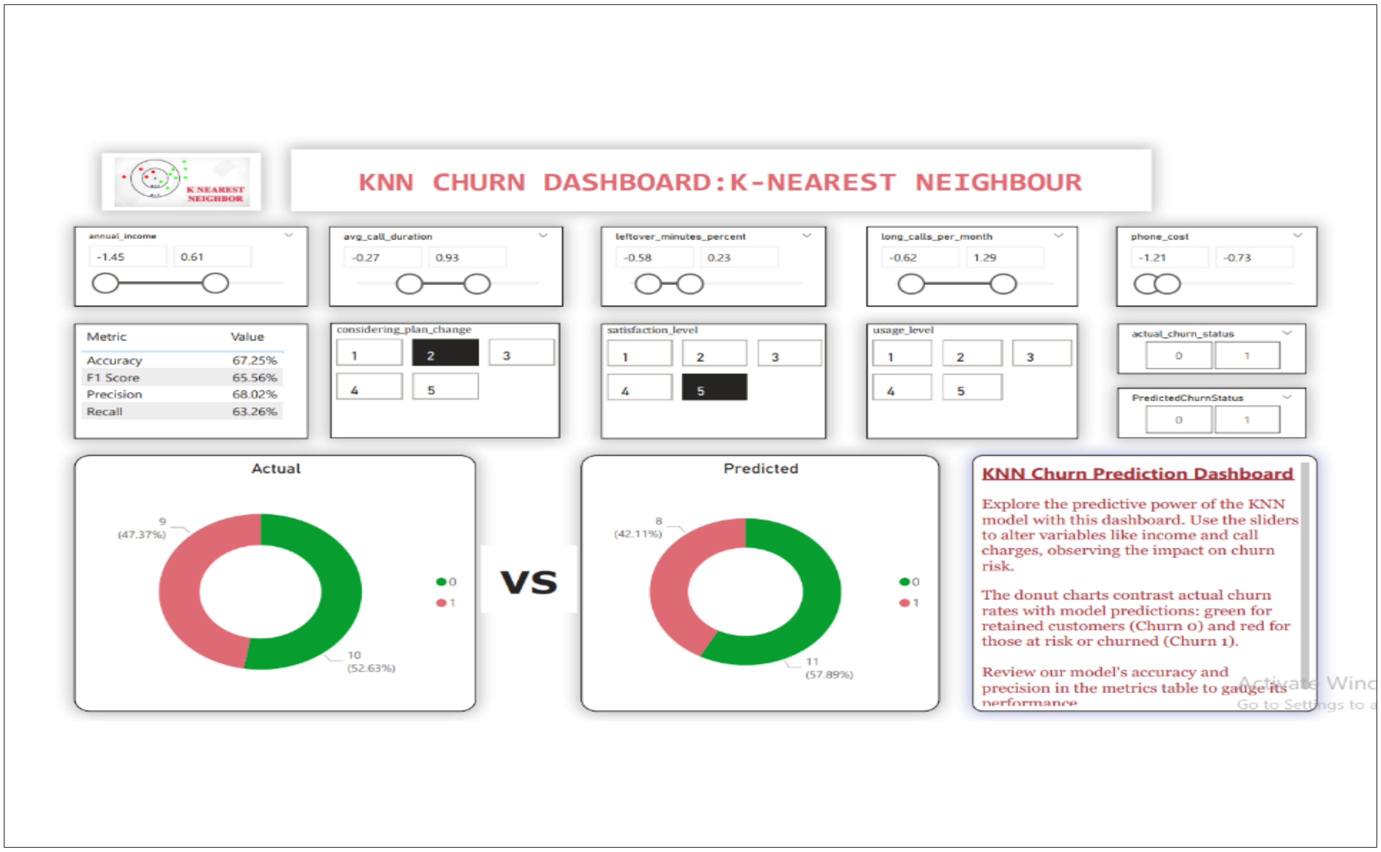
Cluster 2 shows a segment of customers who have high annual incomes and are likely to incur high monthly overcharges. This suggests they are heavy users of the service but may not be on the most cost-efficient plans. Their house values are slightly below average in the data set, maybe they are pragmatic about their spending outside of telecommunications services. The very high phone costs and several long calls per month may mean that they use their phones very often, maybe for business purposes. This is also supported by their longer average call durations.

This group also shows a varied level of satisfaction, meaning there is room for improving customer experience. Their moderate consideration for plan changes indicates a level of contentment but also presents an opportunity for proposing more suitable plans that cater to their heavy usage patterns.

FINAL TASK : PREDICTIVE DASHBOARD

In this final task , the aim is to use the models built to craft a predictive power BI dashboard. The dashboard integrates the decision tree, logistic regression, and kNN model to predict customer churn with the simplicity of a user-friendly interface. Through the use of sliders and input fields in the dashboard, any user can now adjust several predictors like annual income, monthly overcharge, usage level and user satisfaction to instantly see how these changes affect churn probability. This interactive feature not only makes the dashboard practical but also educational. Thus, allowing users to gain insights into the data patterns uncovered.




KNN CHURN DASHBOARD:K-NEAREST NEIGHBOUR
K NEAREST NEIGHBOR
annual_income
-1.45
0.61
avg_call_duration
-0.27
0.93
leftover_minutes_percent
-0.58
0.23
long_calls_per_month
-0.62
1.29
phone_cost
-1.21
-0.73
Metric Value
Accuracy 67.25%
F1 Score 65.56%
Precision 68.02%
Recall 63.26%
considering_plan_change
satisfaction_level
usage_level
actual_churn_status
PredictedChurnStatus
Actual
9 (47.37%)
10 (52.63%)
VS
Predicted
8 (42.11%)
11 (57.89%)
KNN Churn Prediction Dashboard
Explore the predictive power of the KNN model with this dashboard. Use the sliders to alter variables like income and call charges, observing the impact on churn risk.
The donut charts contrast actual churn rates with model predictions: green for retained customers (Churn 0) and red for those at risk or churned (Churn 1).
Review our model's accuracy and precision in the metrics table to gauge its


CUSTOMER CHURN DASHBOARD: LOGISTIC REGRESSION
Logistic Regression
Left Over Minutes
0
15
Annual Income
20007
95612
Monthly Overcharge
-2
194
Avg call Duration
1
7
Long calls P/m
0
18
churn = 1 = 0.84
Total Customers by churn_status
49.26%
50.74%
churn_status
annual_income avg_call_duration considering_plan_change leftover_minutes_percent long_calls_per_month satisfaction_level usage_level monthly_overcharge churn_status
20007 6 no 23 4 very_sat little 36 1
20009 4 actively_looking_into_it 18 20 very_sat very_high 183 1
20012 5 never_thought 9 0 very_unsat little 246 1
20013 1 considering 87 5 unsat avg 75 1
20015 5 no 17 1 unsat avg 0 1
20017 5 never_thought 0 0 very_unsat very_little 0 0
20022 1 no 36 17 avg little 205 1
20024 2 actively_looking_into_it 50 22 very_unsat little 243 1
20028 6 considering 7 20 very_sat very_high 177 0
20029 5 considering 16 11 avg little 206 0

CONCLUSION

This analysis begins from refining data to creating predictive models that turn customer information into actionable insights. I focused on preparing the data meticulously, ensuring data quality, integrity and consistency. Further uncovering hidden trends through Exploratory analysis, and then choosing the right models to meet the specific business needs. From my models built it appears that the decision tree model is more efficient based on the metrics and outcome. However, when choosing the right model fits, it is iterative depending on the goal of the business.

The development of these models is more than just an exercise in data science, it is about gaining a deeper connection with customer behaviours and needs. The resulting dashboard serves as a practical and straightforward interface for complex data-driven predictions. The steps taken in this analysis have practical applications in a real-world business setting and will help make informed decisions, improve customer relations, and evolve the business.

REFERENCES

A. S. Rao, B. V. V. a. H. S., 2021. Role of Exploratory Data Analysis in Data Science. *2021 6th International Conference on Communication and Electronics Systems (ICCES),* Volume 6, pp. 1457-1461.

Blockeel, H. et al., 2023. Decision trees: from efficient prediction to responsible AI. *Frontiers in Artificial Intelligence,* 6(10).

Friedman, J. H., Trevor, H. & Rob, T., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software,* 33(1), pp. 1-22.

Geeks, c., 2023. *What is the Glmnet package in R?.* [Online]
Available at: https://www.geeksforgeeks.org/what-is-the-glmnet-package-in-r/
[Accessed 12 December 2023].

Stoltzfus, J. C., 2011. Logistic regression: a brief primer. *Acad Emerg Med,* 18(10), pp. 1099-104.

Youguo, L. & Haiyan, W., 2012. A Clustering Method Based on K-Means Algorithm. *Physics Procedia* , Volume 25, pp. 1104-1109.

Zhongheng, Z., 2016. Introduction to machine learning: K-nearest neighbors. *Annals of Translational Medicine,* 4(11), pp. 281-218.