Identification of Extended Memory Availability Based on Smartphones Feature

Gözde Nur Özdemir
Department of Statistics
Middle East Technical University
Ankara Turkey
gozdenur104@gmail.com

***Abstract*—This study's main goal is to create Logistic Regression as a base model, SVM, ANN, RF, and XGBoost as machine learning models to categorize extended memory availability (yes/no). This dataset contains many phones features such as price, average rating, 5G capacity or not, processor speed, battery capacity, fast charging availability, RAM capacity, refresh rate and primary camera front.**

Keywords— Exploratory Data Analysis, Statistical Test, Linear Regression, Machine Learning Models.

I. Introduction

Smartphones are essential components in modern life, tools for communication, entertainment and information access. Therefore, smartphones' features are really important for consumers. For example, extended memory availability, which allows increase storage capacity. Several factors, which are price, RAM capacity, processor speed, battery life, camera features and so on, may influence this feature. Understanding these features and their relationships are important both for consumers and manufacturers. In this study, our main research question is: "Which smartphone characteristics significantly predict the availability of extended memory?" To answer this question, we try to find relationship between extended memory availability (classified as yes and no) and other features by using Logistic regression. For example, average rating, battery capacity have positive effect on extended memory availability. However, price and primary camera front have weak negative effect on extended memory availability yes cases. Later, we try XGBoost, Support Vector Machine, Random Forest and Artifical Neural Networks. These models are compared with each other by using confusion matrix.

II. Literature Review

There are a few types of research that analyze smartphones. Mainly, they are related to price variables. One way of deciding whether to buy smartphones or not is price. Reference [1] used 20 input features such as battery capacity, RAM, internal memory, and etc, to accurately classify phones into 4 segments (low, medium, high, very high) by using SVM, RF, decision tree, logistic regression and KNN. These methods were evaluated based on accuracy, precision, recall, and F1-score. SVM achieved 98% accuracy, RF achieved 88.8%, logistic regression achieved 85.5%, KNN achieved 82.6% and decision tree achieved 80.5%. Moreover, they found that when the number of features increases, the price category also increases (e.g., low to medium). Also, there is one more study that is related to average rating which is our sub-research question. Reference [2] showed that user satisfaction is positively associated with performance, display, and material. However, slow charging, price, and poor low-light performance have a negative impact on average rating.

III. Methodology

A. *Dataset*

Reference [3] provided a smartphone dataset that includes many smartphone' features. There are 980 smartphones and 22 variables which are categorical, numerical, and text style. There are 383 NA values in total.

1) *Brand Name, represents the brand name of the phone.*
2) *Model, represents models for each phone.*
3) *Price, is a numeric variable. It is in the respective currency.*
4) *Average Rating, is a score for a phone based on users or reviewers.*
5) *5G or not, 0= no, 1=yes.*
6) *Processor Brand, is a character. It indicates the brand of the processor.*
7) *Number of Cores, has 3 values which are 4,5 and 8*
8) *Processor Speed, is the speed of the processor, measured in GHz.*
9) *Battery Capacity, represents battery size. It was measured in mAh.*
10) *Fast Charging Available, 0=no, 1=yes.*
11) *Fast Charging, is the fast charging capability.*
12) *RAM Capacity, is the random access memory availability measured in GB.*
13) *Internal Memory, is the internal storage capacity of the mobile phone measured in GB.*
14) *Screen Size, is the size of the mobile phone's display screen measured in inches.*
15) *Refresh Rate, shows the screen refresh rate in Hz.*
16) *Extended Memory Availability, 0=no, 1=yes.*
17) *Number of Rear Cameras, is a number of rear-facing cameras.*
18) *Primary Camera Rear, is the resolution of the primary rear camera in megapixels.*
19) *Primary Camera Front, is the resolution of the primary front-facing.*
20) *Resolution Height, is the height dimension of the display screen resolution.*
21) *Resolution Width, is the width dimension of the display screen resolution.*
22) *OS, is the operating system.*

B. *Descriptive Statistics*

To understand the structure and the properties of variables, descriptive statistics were used. For the numeric variables, the minimum, maximum, mean, median, and quantiles were obtained. The mean represents the central tendency, and it is sensitive to outliers. The median represents the central location, and it is not sensitive to outliers. A summary of descriptive statistics of numeric variables is shown below.

TABLE I. Summary of Numerical Variables

Statistic	Camera Rear	Camera Front	Resolution Height	Resolution Width	Log Price	Fast Charging
Min	2	0	480	480	8.16	10
1st Qu.	24	8	1612	1080	9.473	18
Median	50	16	2400	1080	9.903	33
Mean	50.32	16.59	2215	1076	10.032	46.3
3rd Qu.	64	16	2408	1080	10.477	66
Max	200	60	3840	2460	13.385	240
NA	200	5	0	0	0	0

TABLE II. Summary of Numerical Variables

Statistic	Average Rating	Processor Speed	Battery Capacity	RAM Capacity	Screen Size	Refresh Rate
Min	6	1.2	1821	1	3.54	60
1st Qu.	7.4	2.05	4500	4	6.5	60
Median	8	2.3	5000	6	6.58	90
Mean	7.826	2.427	4818	6.56	6.537	92.26
3rd Qu.	8.4	2.84	5000	8	6.67	120
Max	8.9	3.22	22000	18	8.03	240
NA	101	42	11	0	0	0

The primary camera rear ranges between 2 to 200, and it has 200 missing values. The primary camera front's mean and median are very close to each other so it is distributed symmetrical. It has 5 missing values. Resolution width ranges between 480 to 2460. To make a symmetric distribution, a log of price variables was used. It is mean and median are close to each other, and it ranges between 8.16 to 13.385. The average rating seems symmetric, and it has 101 missing values. Processor speed ranges between 1.2 to 3.22. Battery capacity has 11 missing values. Screen size has a median of 6.58 and a mean of 6.537, so it is symmetric. The refresh rate has a median of 90 and a mean of 92.26.

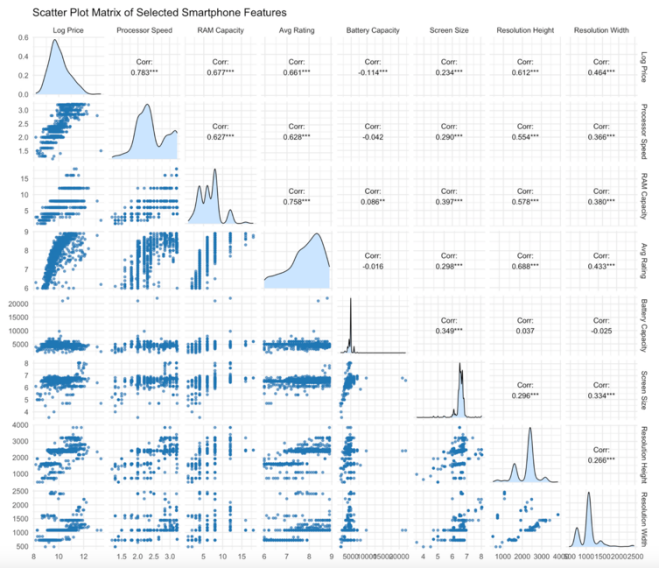

Fig. 1. Stacked Bar Chart

The scatter plot matrix visualizes the pairwise relationships between numerical features. For this plot we did not use all features, we used some selected features based on their correlation strength. There is a strong and approximately linear relationship between log price and several features such as processor speed (r = 0.78), RAM capacity (r = 0.68), and average rating (r = 0.66). Also, there is a relationship between average rating and resolution height (r= 0.668).

Descriptive statistics for the binary and categorical variables are presented below. From these tables, distributions of different categories within the dataset can be seen.

TABLE III. Summary of Binary Variables

Levels	Fast Charging	5G or not	Extended Memory
0	143	431	362
1	837	549	618

Our dataset contains 143 cases where fast charging is not available and 837 cases where it is available. There are 549 phones with 5G support and 431 not 5G support. Additionally, 618 phones have extended memory available, while 362 do not have extended memory available.

TABLE IV. Summary of Categorical Variables

Levels	Number of Cores	Levels	Number of Rear Cameras	Levels	Internal Memory
4	36	1	65	16	12
6	39	2	208	32	67
8	899	3	551	64	193
NA	6	4	156	128	523
				256	157
				512	22

The majority of phones have 8 number of cores which 899 of our phones fall into this category. The number of rear cameras' has mainly 3 and 4 levels. We have 6 levels for internal memory, these are 16, 32, 64, 128, 256, and 512. The most common value is 128 GB, observed in 523 devices. The least common value is 16 GB, observed in 12 devices.

C. Exploratory and Confirmatory Data Analyses

In this section, 3 different research questions are addressed to explore the structure of the dataset and highlight the critical findings that directly inform the modeling process. For the EDA part complete dataset was used without filling missing values. For the CDA part imputed dataset was used. Imputation method and strategy will be explained after this part.

1) Do devices with higher number of cores tend to support 5G more frequently?

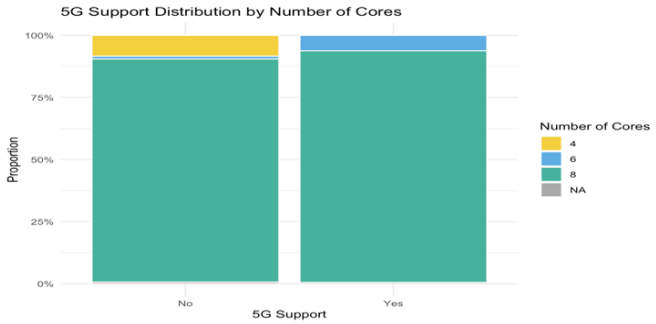

Fig. 2. Stacked Bar Chart

The stacked bar chart illustrates the relationship between the number of cores and 5G availability (yes/no). Phones with more cores tend to support 5G more frequently than those with fewer cores. The chart shows that devices with 8-core support for 5G are more prevalent than those without support. Also, there is no 5G support for 4-core devices. 4 core devices do not support 5G.

To decide if a higher number of cores tends to support 5G more frequently or not, the Fisher Exact test with Simulation was used since some cells expected frequencies are smaller than 5 and our table is bigger than 2 by 2.

TABLE V. Summary of Fisher Test

Statistic	Result
p-value	0.00001
Replicates	100000

Each observation in the dataset corresponds to different smartphones so we confirmed that the independence assumption was met. The *p-value* is 0.00001, which is significantly below the significance level of 0.05. This result provides that the number of processor cores and the availability of 5G support are significantly associated.

2) Are there any significant effects between categorical variables and average rating?

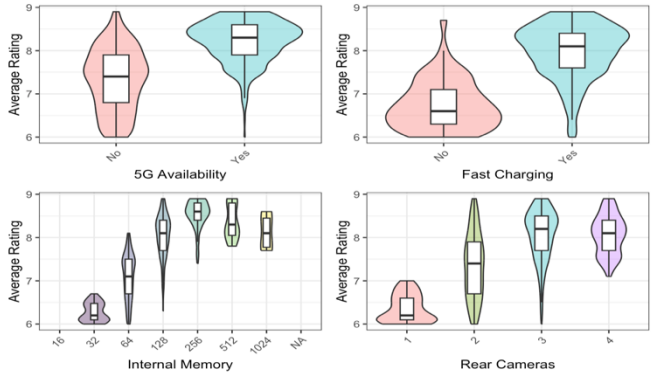

Fig. 3. Violin Plots

Phones with 5G support have higher average ratings compared to those without. The 5G no category shows more variability than the 5G yes category. Moreover, phones with fast charging capabilities are rated much higher than those without. Higher internal memory devices have higher ratings, such as 1024, 512, and 256 GB. However, low internal memory categories such as 32 and 64 show lower ratings. For the 16 GB internal memory category, there are no average rating observations, so it creates an empty area. Finally, as the number of rear cameras increases, the average rating also increases. The rating distributions for devices with 2 and 3 rear cameras are wider compared to those with 1 or 4 cameras.

TABLE VI. Summary of Fisher Test

Variable	Test Used	Statistic	df	p-value
Average Rating	Shapiro-Wilk	W = 0.925	—	< 0.05
5G	Wilcoxon Rank Sum	W= 30839	—	< 0.05
Fast Charging	Wilcoxon Rank Sum	W = 8525	—	< 0.05
Internal Memory	Kruskal-Wallis	Chi-sq= 600.08	df = 7	< 0.05
Rear Cameras	Kruskal-Wallis	Chi-sq= 308.96	df = 3	< 0.05

The first normality assumption was checked. The *p-value* is smaller than the significant level, so not normally distributed. Also, observations are independent as each row is a unique smartphone. Non-parametric methods were used to assess the significance of categorical variables about average rating. For 5G and fast charging availability variables, we used the Wilcoxon test to assess whether the distributions of average ratings differ between the groups. For the internal memory and the number of rear cameras, we used the Kruskal-Wallis test. All comparison seems statistically significant since p values are smaller than 0.05. So, we can say that there are statistically significant differences in average rating depending on several features. Fast Charging availability, and 5G availability are binary variables. For these variables, post-hoc testing was not needed. For the non-binary variables, we need to analysis using Dunn's test to identify which specific groups differed.

TABLE VII. Summary of Dunn's Test

Comparison of Internal Memory Groups		
Comparison	P.adj	P < 0.05?
128 - 16	≈0	< 0.05
128 - 256	≈0	< 0.05
16 - 256	≈0	< 0.05
128 - 32	≈0	< 0.05
256 - 32	≈0	< 0.05
16 - 512	≈0	< 0.05
32 - 512	≈0	< 0.05
128 - 64	≈0	< 0.05
256 - 64	≈0	< 0.05
512 - 64	≈0	< 0.05
1024 - 32	0,00013	< 0.05
32 - 64	0,00014	< 0.05
128 - 512	0,00019	< 0.05
1024 - 16	0,00111	< 0.05
1024 - 64	0,00698	< 0.05
256 - 8	0,01409	< 0.05
512 - 8	0,01538	< 0.05
16 - 64	0,10947	> 0.05
1024 - 8	0,10947	> 0.05
128 - 8	0,10947	> 0.05
1024 - 256	0,14577	> 0.05
1024 - 512	0,18117	> 0.05
64 - 8	0,65484	> 0.05
1024 - 128	0,90778	> 0.05
16 - 32	0,99506	> 0.05
16 - 8	0,99762	> 0.05
32 - 8	0,99762	> 0.05
256 - 512	0,99803	> 0.05

Comparison of Rear Camera Groups		
Comparison	P.adj	P < 0.05?
1 - 2	≈0	< 0.05
1 - 3	≈0	< 0.05
2 - 3	≈0	< 0.05
1 - 4	≈0	< 0.05
2 - 4	≈0	< 0.05
3 - 4	0,21809	> 0.05

According to Dunn's post-hoc test results, there are many statistically significant differences in internal memory groups based on average ratings. For example, devices with 128 GB of internal memory were significantly different from those with 16 GB, 32 GB, 64 GB, and 512 GB. However, no statistically significant difference was observed between groups such as 16 GB vs. 32 GB (p= 0.99506), 1024 GB vs. 256 GB (p = 0.145), and 256 GB vs. 512 GB (p = 0.998).

Also the rear camera comparison revealed statistically significant differences between several groups. For example, devices with 1, 2, and 3 cameras were all significantly different from each other. However, there was no significant difference in average rating between devices with 3 and 4 rear cameras ($p = 0.218$).

3) Is there a significant relationship between binary features and extended memory?

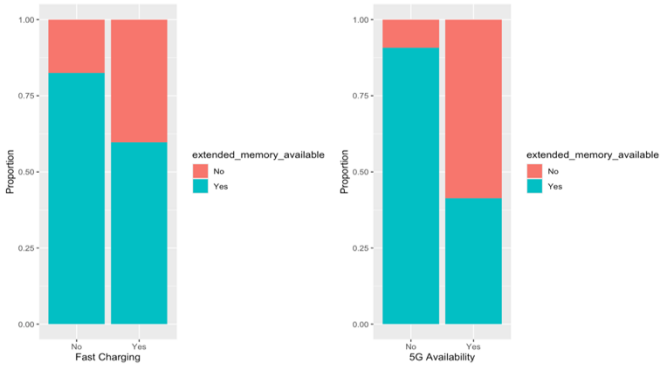

FIG. 4. STACKED BAR CHART

Our main interest is which factors have an effect on extended memory. To get an idea before the modeling part we looked at extended memory vs binary variables association. Fig. 4 shows fast charging no phones have more extended memory. Also, 5G availability shows a negative association with extended memory variables. For this analysis, we used the Chi-squared test since the expected cell frequencies were all greater than 5 and the observations were assumed to be independent.

TABLE VIII. Chi- Squared Tests

Test	Chi-squared	df	*p-value*
Fast ChargingAvailability vs Extended Memory	26.24	1	$p < 0.05$
5G Availability vs Extended Memory	250.54	1	$p < 0.05$

TABLE VIII illustrates that fast charging availability and extended memory have a statistically significant association which yields a chi-squared value of 26.24. Moreover, 5G availability and extended memory also have statistically significant associations which yields a chi-squared value of 250.54. To examine the importance of these features and their association with extended memory, logistic regression is used.

D. Influential Observations

Before deciding final logistic regression model first influential points detection was made based on standard diagnostic measures: Cook's Distance, leverage values, and studentized residuals. Observations were flagged as influential if they exceeded one or more of the following thresholds [4] : Cook's Distance > 4/n, hat values > 2p/n, or studentized residuals > 2. These criteria were derived from the diagnostic framework introduced by Belsley, Kuh, and Welsch [4] After filtering out these points, full and without influential logistic regression was built.

TABLE IX. Model Evaluation Before and After Influential Point Removal

Metric	Full Model	Influentials Removed Model
Accuracy	0.8969	0.964
95% CI	(0.8762, 0.9153)	(0.949, 0.9756)
NIR	0.6306	0.6487
Kappa	0.7771	0.9206
McNemar's P Value	0.2325	0.2012
Sensitivity	0.9288	0.9797
Specificity	0.8425	0.9352
Pos Pred Value	0.9097	0.9654
Neg Pred Value	0.8739	0.9614
Prevalence	0.6306	0.6487
Detection Rate	0.5857	0.6355
Detection Prevalence	0.6439	0.6583
Balanced Accuracy	0.8857	0.9574

TABLE IX shows us cleaned dataset achieved higher accuracy (0.964), and improved balanced accuracy (0.9574). This result shows us that removing influential points are good choice for this situation.

E. Missingness

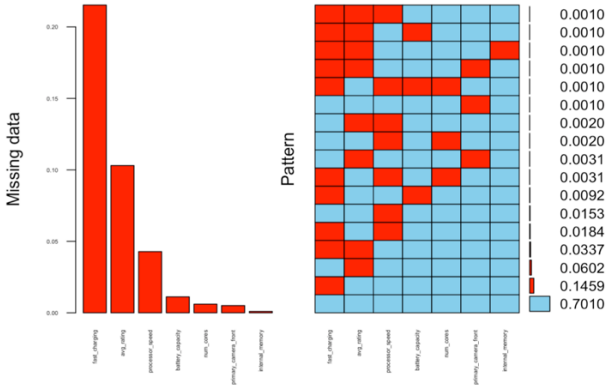

FIG. 5. MISSING VALUE PATTERN AND PROPORTION PLOT

Fig. 5 shows us, that the dataset contains missing values. The most notable variables fast charging, average rating, and processor speed, with fast charging having over 20% missingness values. We need to make an imputation for these missing values, so we used the MICE method, which is suitable for datasets with both continuous and categorical variables.

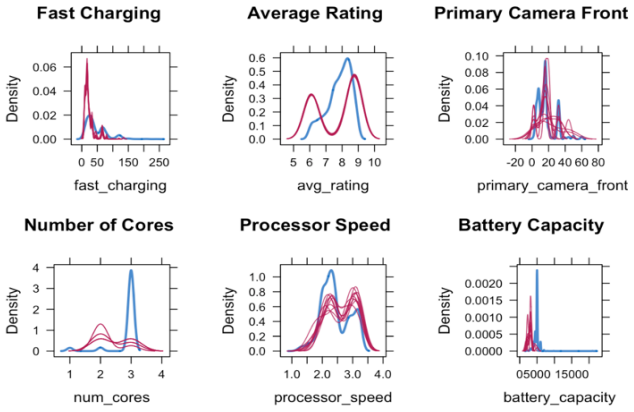

Fig. 6. Density Plots of Imputed vs. Observed Distributions

Overall, imputation methods seem nice for each of the variables since imputed and non-imputed density curves look similar for each case. So, our imputation seems valid.

F. *Modelling*

1. Logistic Regression

To model the extended memory availability, we first fitted logistic regression as a base model. First, all predictors such as price, processor speed, number of cores, internal memory, and more were included in our model, later insignificant ones removed from our analysis. To further validate feature selection, LASSO regularization was applied. This approach confirmed that the reduced model is robust and suitable for our analysis. Before interpreting the model, we performed assumption checks.

TABLE X. VIF Values

Variable	VIF
Price	1.8313
Averagr Rating	4.4553
5G	1.4422
Processor Speed	1.7758
Battery Capacity	1.6075
Fast Charging	1.2618
RAM Capacity	2.2705
Refresh Rate	1.5899
Primary Camera Front	1.5595

There is no multicollinearity problem since all of the VIF values are smaller than 5.

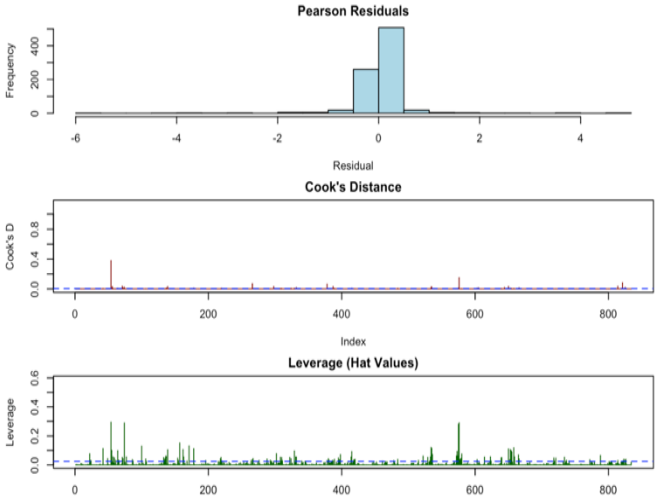

FIG. 7. DIAGNOSTIC PLOTS: RESIDUALS, COOK'S D, AND LEVERAGE

The top panel shows us residuals are tightly clustered around zero and seem approximately symmetrically distributed. The middle panel shows us none of the data points surpass the influence threshold's limit, so our model seems stable across the dataset. The bottom panel shows us no single observation disproportionately affected the estimation of coefficients.

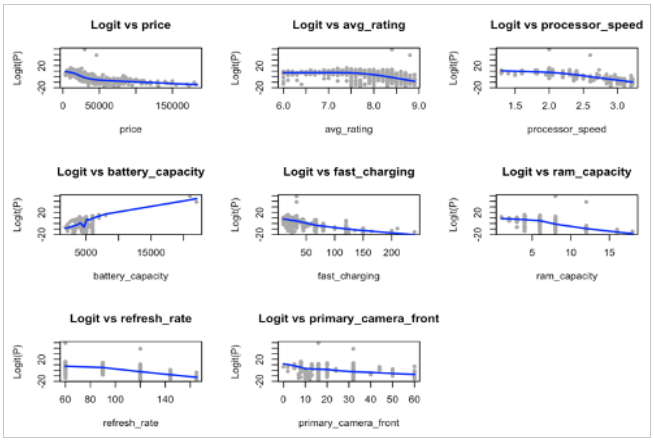

Fig. 8. Logit Linearity Plots

The blue LOESS curves across all variables demonstrate approximately linear relationships with the logit of the outcome. Although slight curvatures were observed in variables such as average rating and price, they do not represent serious violations of the linearity assumption. Therefore, the assumption of linearity in the logit is reasonably satisfied.

The final logistic regression model is expressed as:

$$\begin{aligned} logit(P) = &-15.404 \times Intercept \\ &- 0.00005527 \times Price \\ &+ 4.62783 \times Average\ Rating \\ &- 3.80527 \times 5G\ Yes \\ &- 6.29384 \times Processor\ Speed \\ &+ 0.00243 \times Battery\ Capacity \\ &- 0.04961 \times Fast\ Charging \\ &- 0.83189 \times RAM\ Capacity \\ &- 0.04078 \times Refresh\ Rate \\ &- 0.024 \times Primary\ Camera \end{aligned}$$

Null deviance: 1081.30 on 833 degrees of freedom
Residual deviance: 156.14 on 824 degrees of freedom
AIC: 176.14

The average rating exhibits the most substantial positive effect on the log odds of 4.628, suggesting that higher customer ratings are associated with extended memory. However, processor speed, 5G availability, and RAM capacity have negative effects on extended memory. Same result was obtained from Section III-C, under RQ 3. The model performs well; there is a significant reduction from null deviance to residual deviance.

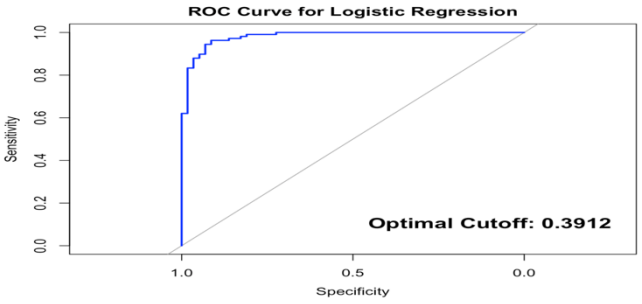

Fig. 9. ROC Curve

The area under the curve seems close to 1, so our model well classifies. The optimal cutoff point was 0.3912.

TABLE XI. Train vs Test Performance

Metric	Train	Test
Accuracy	0.9656	0.9337
Kappa	0.9243	0.8548
Sensitivity	0.9447	0.9138
Specificity	0.9769	0.9444
Precision (PPV)	0.9569	0.8983
NPV	0.9702	0.9533
Balanced Accuracy	0.9608	0.9291

The performance metrics such as accuracy, sensitivity, specificty, and accuracy are consistently high for both train and test sets. So, model generalizes well to unseen data. The model shows no signs of overfitting or underfitting.

2. *Support Vector Machine*

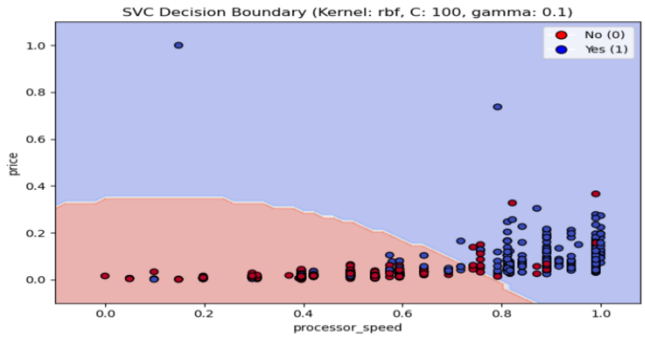

Fig. 10. SVM Decision Boundary

We implemented a Support Vector Machine (SVM) model and conducted hyperparameter tuning to enhance its predictive performance. A grid search was performed over various kernel types (linear, polynomial, RBF, and sigmoid), regularization strengths (C = 0.1, 1, 10, 100), and gamma values (0.01, 0.1, 1). As shown in Fig. 10, the optimal configuration was found to be RBF kernel with C = 100 and gamma = 0.1. The tuned model achieved a test accuracy of 0.852 and an F1-score of 0.89, with high recall (0.95) and satisfactory precision (approximately 0.84).

3. *Artificial Neural Networks*

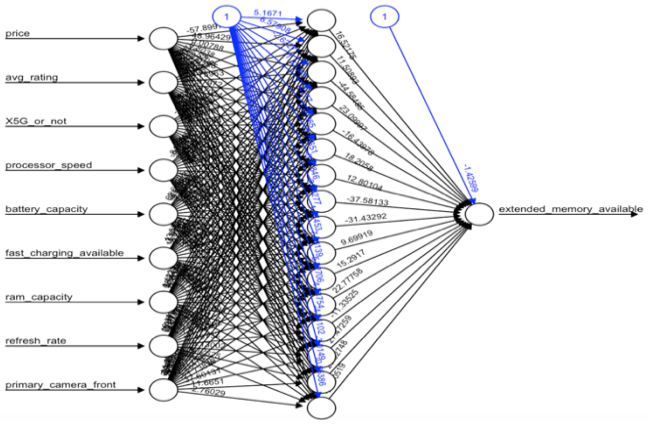

Fig. 11. Neural Network Architecture

ANN was inspired by the structure of the human brain. We applied comprehensive hyperparameter tuning to identify optimal parameters such as activation functions (relu, tanh, elu, selu), dropout rates (0.1, 0.2, 0.3, 0.4), and batch size (16, 32, 64). Eearly stopping was used to prevent overfitting. The optimal ANN configuration was found to have a single hidden layer with 16 neurons, relu activation function, dropout rate of 0.1, and a batch size of 16. The model achieved an accuracy of 0.8622, a precision of 0.8702, a recall of 0.9194, and an F1 score of 0.8941.

Fig. 11 presents the visualization of the best-performing ANN model. The diagram displays the input features, one hidden layer, and the output node.

4. *Random Forests*

RF model was optimized by using GridSearchCV with number of trees (50, 100, 150), maximum depth (3, 5, None), minimum samples to split (2, 3, 4), minimum samples leaf (1, 2, 4) and maximum features ('sqrt', 'log2', None). The best combination selected by maximum depth was 5, maximum feature was "sqrt", minimum samples leaf was 1, minimum samples split was 3, and number of trees as 150. The final model achieved accuracy of 0.852 precision of 0.8682, recall of 0.9032 and F1 score of 0.8854.

5. *XGBoost, LightGBM and CatBoost*

We used 3 popular algorithms: XGBoost, LightGBM, and CatBoost. Optuna was used for only these 3 models in our study because CatBoost has several complex parameters such as colsample_bylevel, reg_lambda, and depth. So, manual grid search could be inefficient.

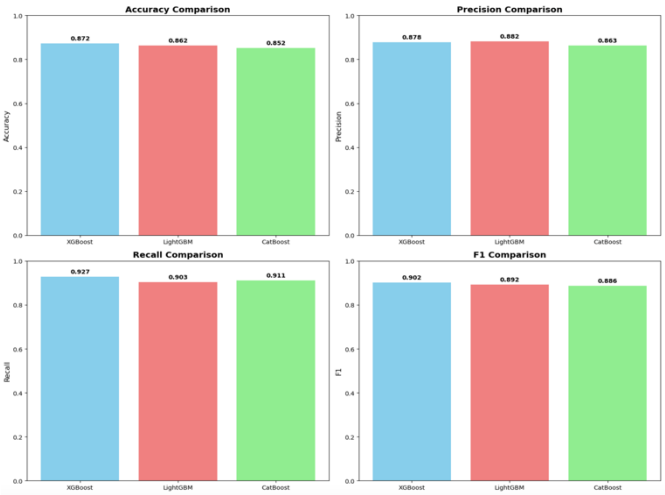

Fig. 12. Comparison of Three Boosting Models

XGBoost has the highest F1 score (0.902), recall (0.9274) and accuracy (0.8724). For the precision LightGBM gives the best result (0.882). XGBoost achieved the best overall performance.

6. *Model Comparison*

TABLE XII. Performance Metrics on Test Dataset

Model	Accuracy	F1 Score	Precision	Recall
SVM	0.852	0.8906	0.8369	0.9516
ANN	0.8622	0.8941	0.8702	0.9194
Random Forest	0.852	0.8854	0.8682	0.9032
XGBoost	0.8724	0.902	0.8779	0.9274

The evaluation metrics include accuracy, F1 score, precision, and recall. SVM achieved the highest recall (0.9516); however, it had the lowest precision (0.8369). ANN achieved a balance between precision and recall, with a strong F1 score (0.8941). RF was performed with a precision of 0.8682 and a recall of 0.9032. XGBoost emerged as the best-performing model. It achieved the highest accuracy, F1 score, and precision. Considering all evaluation criteria, XGBoost is selected as the best model.

TABLE XIII. Performance Metrics on Train Dataset

Model	Accuracy	F1 Score	Precision	Recall
SVM	0.9286	0.9202	0.8858	0.9575
ANN	0.8941	0.9177	0.899	0.9372
Random Forest	0.9286	0.9441	0.9311	0.9575
XGBoost	0.9401	0.953	0.9426	0.9636

We also looked at train performances to see if there were any underfitting and overfitting problems. Across all models, no moderate or severe overfitting or underfitting was observed. The train and test scores remained relatively consistent across all models.

7. *Final Model*

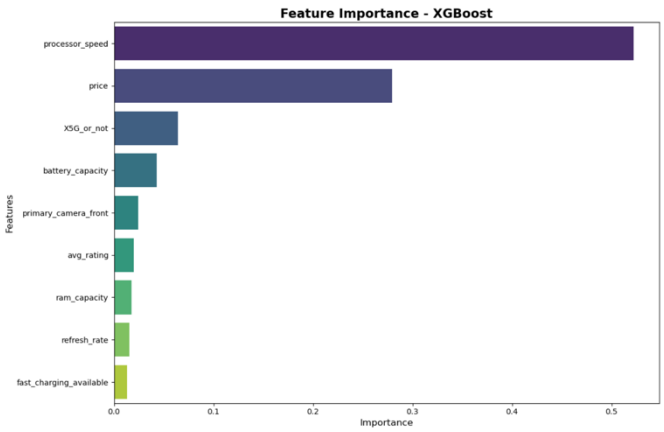

Fig. 13. Feature Importance Plot of the XGBoost Model

Fig. 13 provides information about the variable's importance in predicting whether a smartphone has extended memory availability. The processor speed is the most influential feature. The price is the second most important feature. However, fast charging, refresh rate, and RAM capacity play less important roles in deciding extended memory classification.

IV. Conclusion

The first part of the study was aimed at understanding data structure and related features. In this part, our main goal is to find if there is a relationship between extended memory availability and other features. We saw that extended memory availability was negatively associated with both fast charging and 5G availability. Also, we found that there is a very strong positive correlation between average rating and log price, RAM capacity and average rating, and RAM capacity and log price. Later, influential points detection and missing values imputation were made by using the MICE method. The second part of our analysis, models were built to determine extended memory availability conditions. As a base model, logistic regression was used. Logistic regression performs well; it has high accuracy, sensitivity, specificity, and F1 score. Moreover, for the sake of the study, we also build machine learning models. For the machine learning models XGBoost, CatBoost, LightGBM, random forest, support vector machine, and artificial neural network were used. Based on their test performances, XGBoost was chosen as the best model. It achieves the best accuracy, F1-score, and precision values. The most influential factor in predicting extended memory availability was processor speed, followed by price. However, features such as fast charging availability, refresh rate, and RAM capacity have minimal predictive power.

Marketing teams can use these analyses to make recommendations to decision-makers and conduct further investigations. For example, when advertising a phone with extended memory, highlight its fast processor and price. Also, different customer groups may want different features; for example, gamers will care more about fast processors and big memory. Developing teams can use these models' results to design smartphone features accordingly.

For further analysis, our models rely on structured technical features. Incorporating external, unstructured features may enrich the model. NLP techniques may be used to extract features from unstructured text data. Moreover, the model performs well on historical data; deploying it in a real-world setting also may improve our model's performance.
In summary, our main goal is to predict extended memory availability based on smartphone features. So base logistic regression and machine learning models were built to predict more effectively.

References

[1] N. Sunariya, A. Singh, M. Alam, and V. Gaur, “Classification of mobile price using machine learning,” in Proc. Int. Conf. Intelligent Computing and Smart Communication, New Delhi, India, Mar. 2024, pp. 1–6.

[2] A. Jabeen, M. Shahbaz, and N. Asghar, “A deep learning approach to analyze online reviews for smartphone features and user satisfaction,” in Proc. 3rd Int. Conf. Data Science and Information Technology (DSIT), Shanghai, China, 2024, pp. 156–162.

[3] M. Baloch, “Smartphone dataset,” *Kaggle*, 2023. [Online]. Available: https://www.kaggle.com/datasets/muzammilbaloch/smartphone-dataset

[4] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York, NY, USA: Wiley, 1980.