Team 1:

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

ANALYSIS ON Google play store apps

1. Snehitha Tadapaneni
2. Sai Rachana Kandikattu
3. Amrutha Jayachandradhara
4. Pramod Krishnachari
5. Wilona Nguyen

Problem Statement

The objective of our project is to develop and evaluate the best classification model to predict if a given application is succesful or not based on a range of festures.

Overview

- Introduction
- About the Dataset
- SMART Question
- Evaluation
- Conclusion

About the Dataset

The dataset used in this study is a popular collection of Google Play Store apps, sourced from Kaggle . It contains detailed information about 10,841 apps, organized across 13 variables, with each row representing an individual app from 2010 - 2018.

KAGGLE

Features in the Dataset

- App
- Category
- Rating
- Reviews
- Size
- Installs
- Type
- Price
- Content Rating
- Genres
- Last Updated
- Current Version
- Android Version

Data Cleaning

1

Dropped Duplicated Apps

{APPS}

2

Converted Data Type to INT

{Price, Installs, Reviews}

3

Dropped Missing Values

{PRICE}

4

Replaced missing values with mean

{SIZE}

5

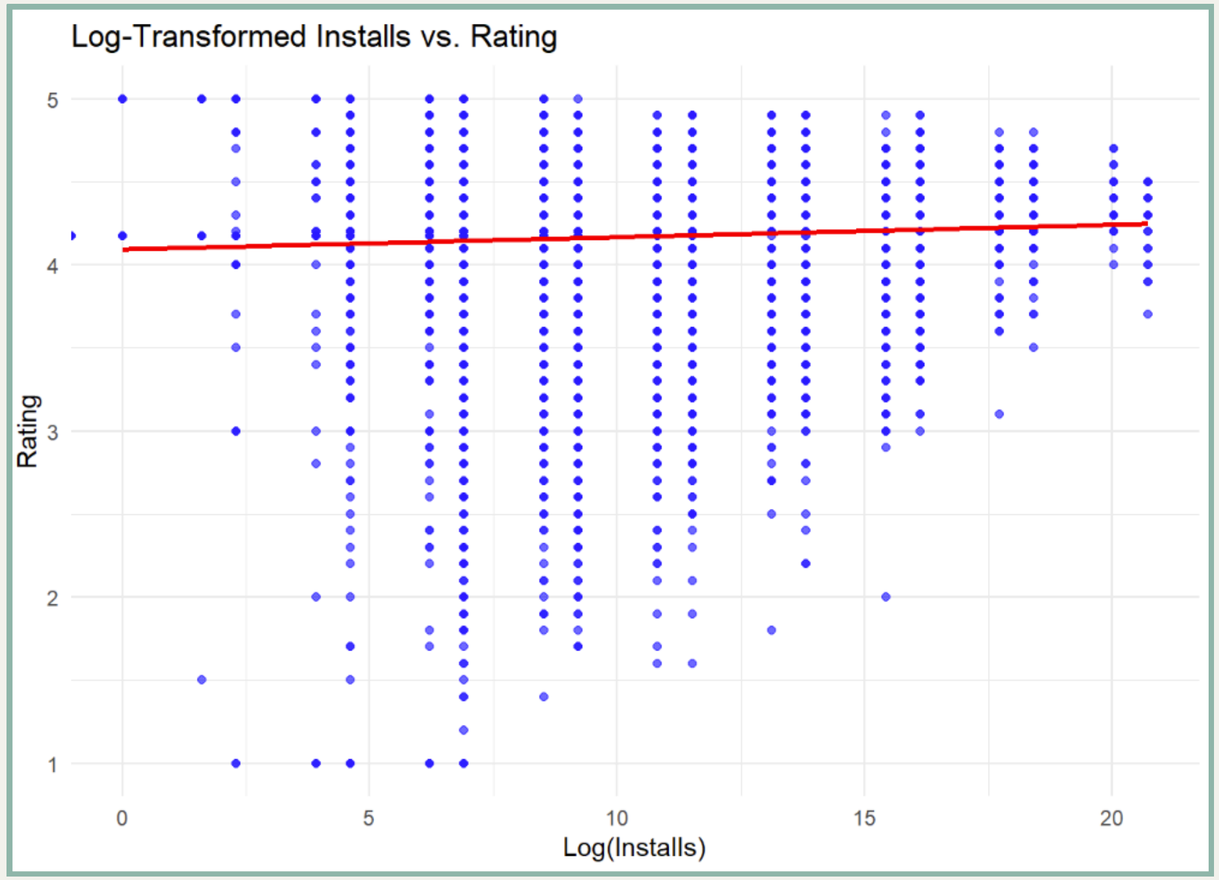
Dropped Irrelevant Columns

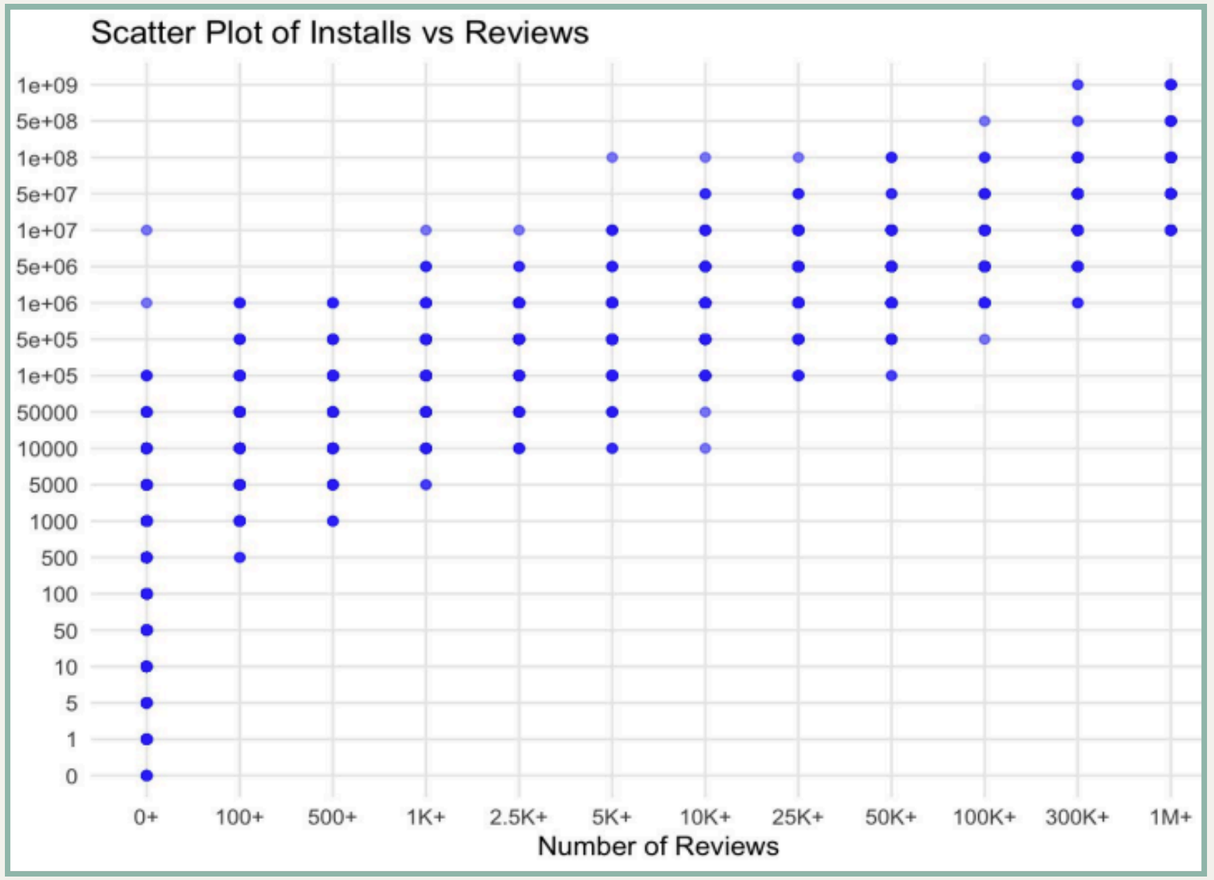
{Type, Android&Current Version, Genre}

Exploratory Data Analysis

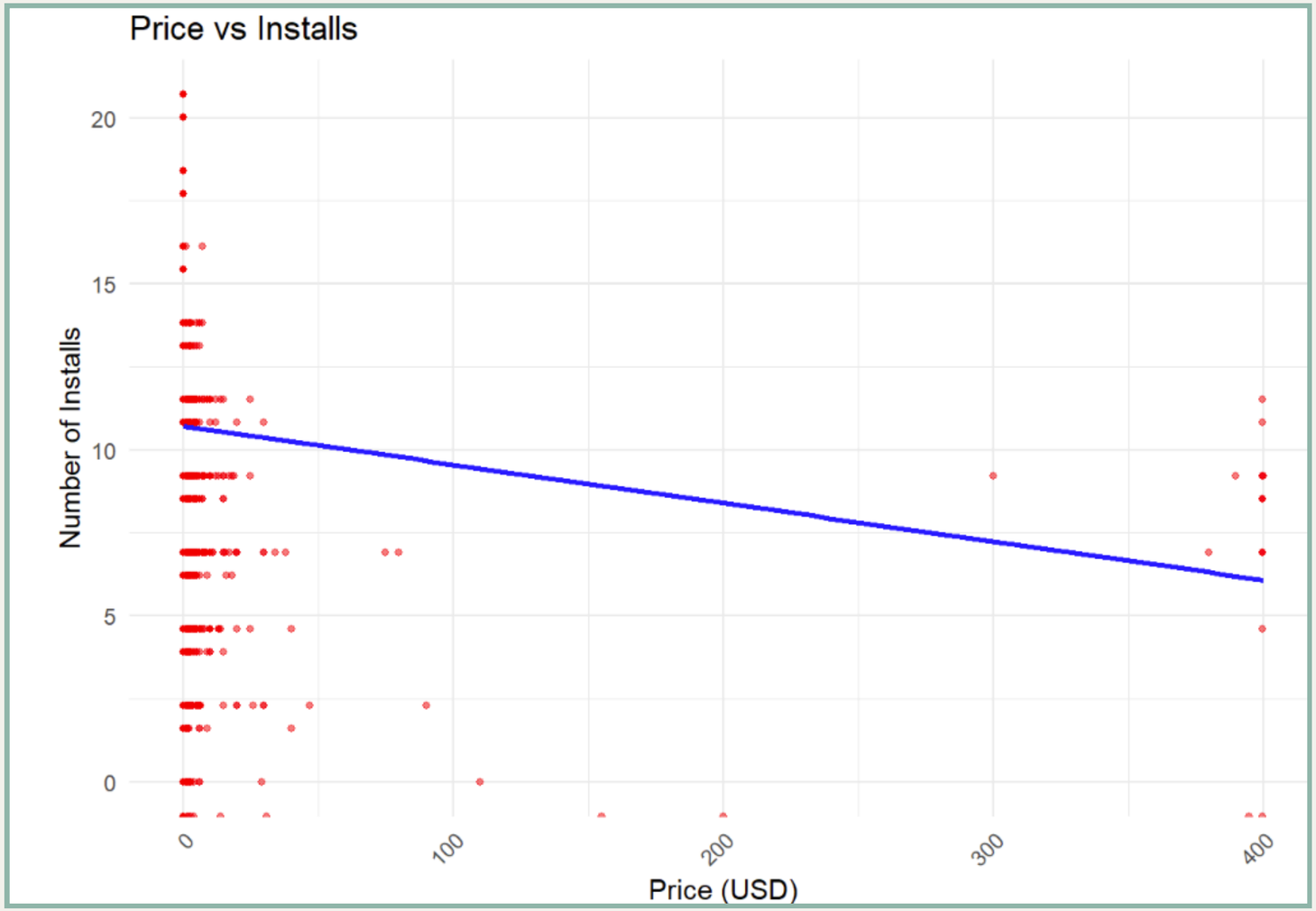
Our target variable is ‘Installs’.
The number of installs can determine if the App is a SUCCESS or NOT A SUCCESS.

Installs vs Reviews & Ratings

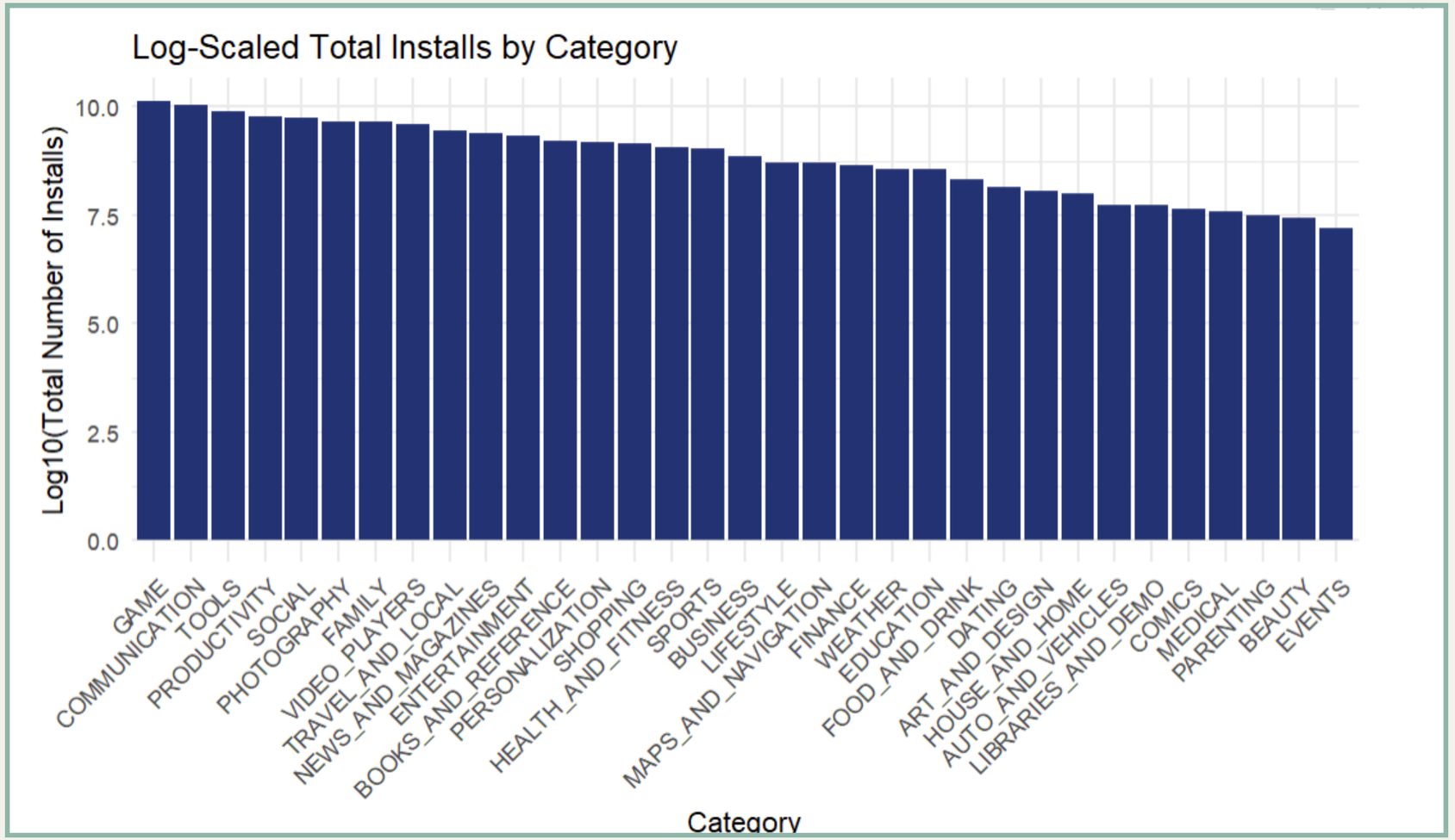




Installs vs Price



Installs vs Categories



Data Preprocessing

Installs	→	Number of Installs	→	Categories: “Low” (0) or “High” (1)
Last Updated	→	Days Since Last Update	→	Relative days
Categories	→	Categories	→	One hot encoding to list of categories

SMART QUESTION

Which are the top 5 app categories, as identified by classification models (logistic regression, SVM, XGBoost, KNN, and random forest), that significantly influenced app success (measured by installs) based on app data from 2010 to 2018, and how accurately can these models predict success trends within this time period?

1
Logistic Regression

2
K- Nearest Neighbour

3
Random Forest

4
Gradient Boost Classifier

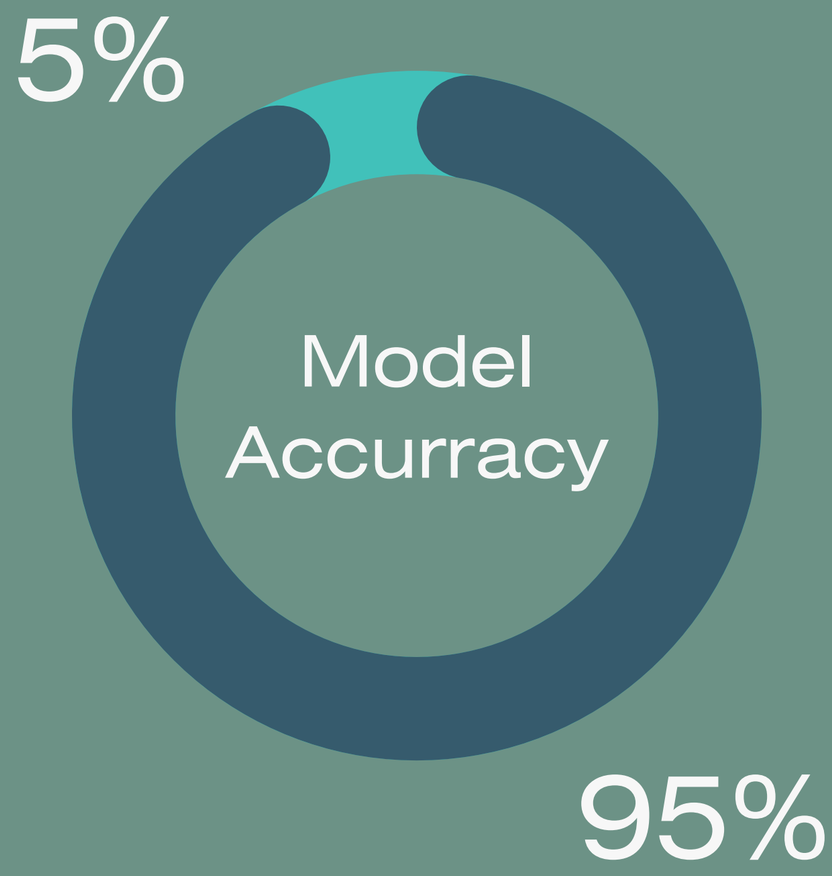
5
Support Vector Machine

Logistic Regression

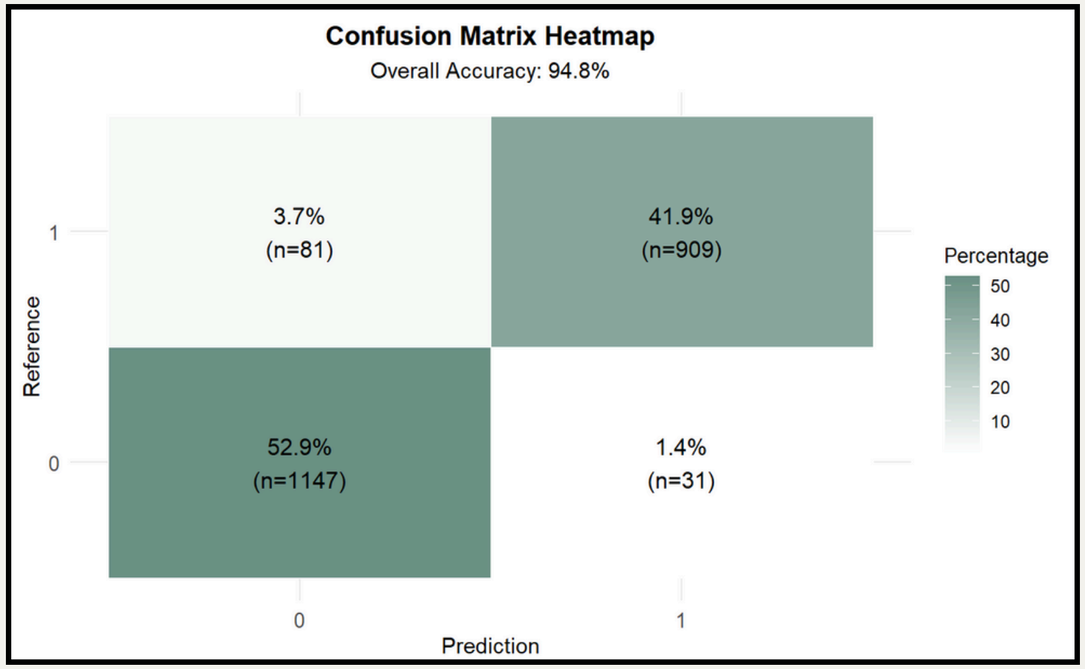
Model Development:

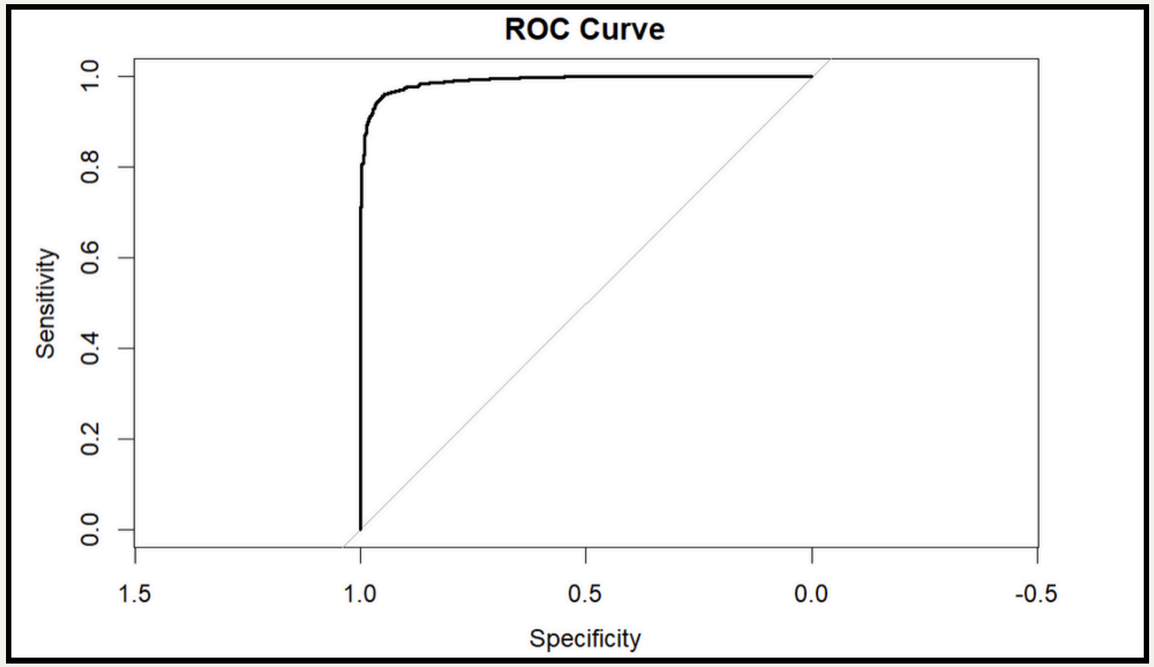
- Stepwise Selection with the lowest AIC
- Number of Features Selected: 16 out of 39

(Rating, Reviews, Price, Content.Rating, Last.Updated, catBUSINESS, catCOMICS, catEDUCATION, catEVENTS, catFINANCE, catHOUSE_AND_HOME, catMAPS_AND_NAVIGATION, catMEDICAL, catNEWS_AND_MAGAZINES, catPHOTOGRAPHY, catSPORTS)



Model’s Performance Overview:

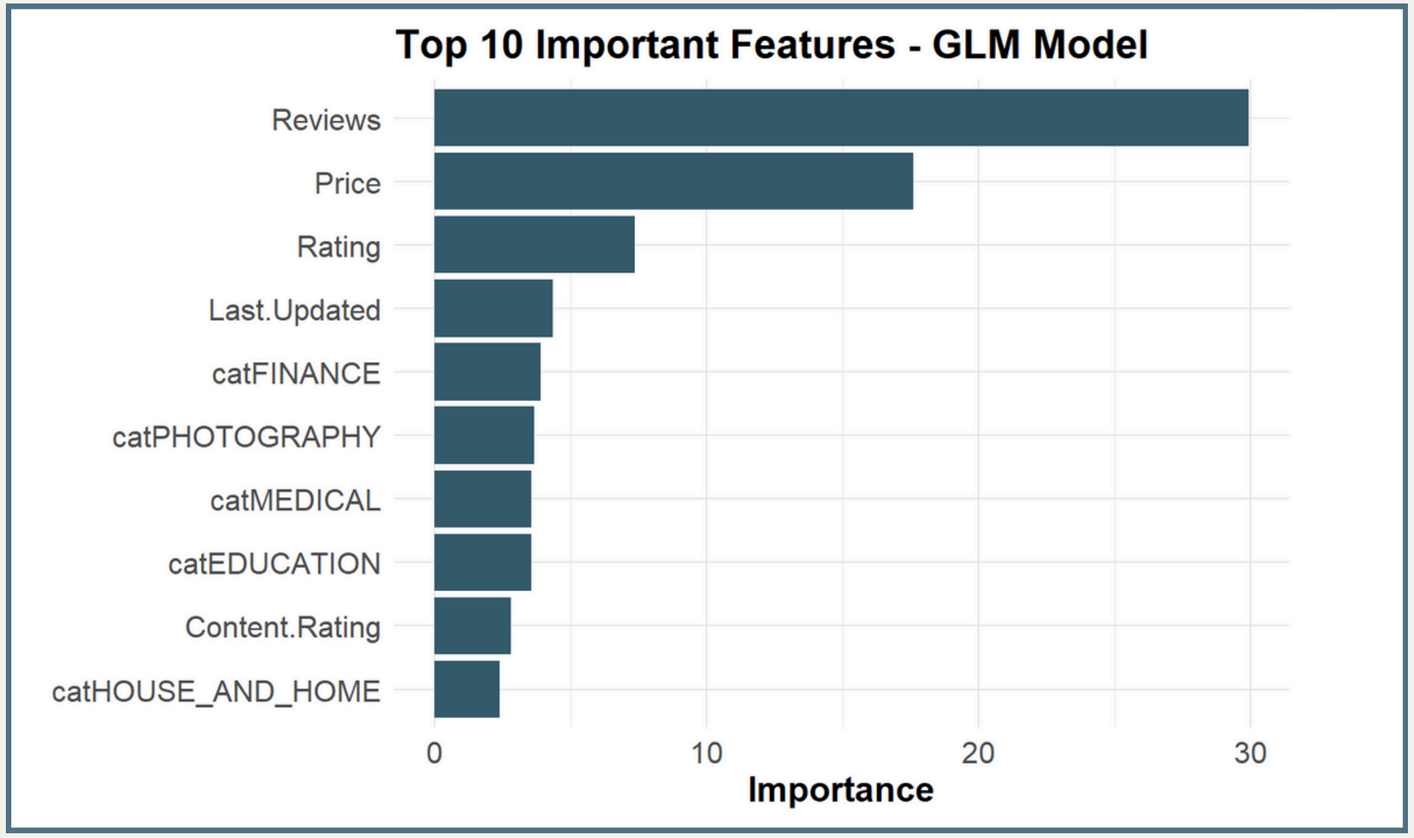




Metric	*Percentage*
Accuracy	0.948
Precision	0.934
F1 - Score	0.953
Recall	0.973

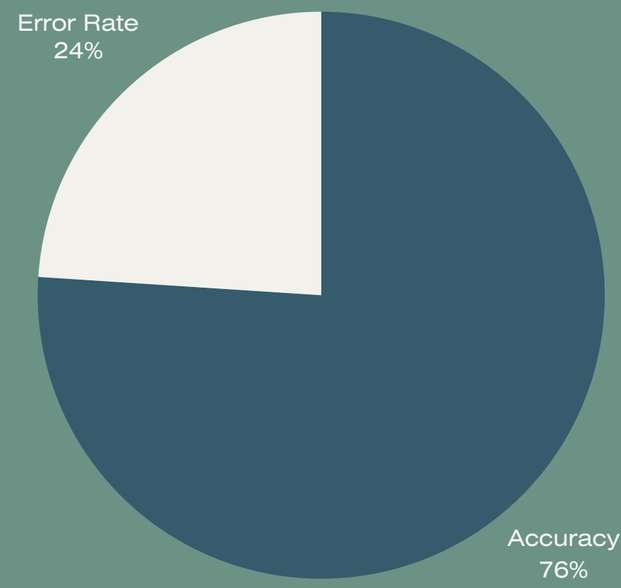
AUC Score: 0.9905884

Feature Importance Plot



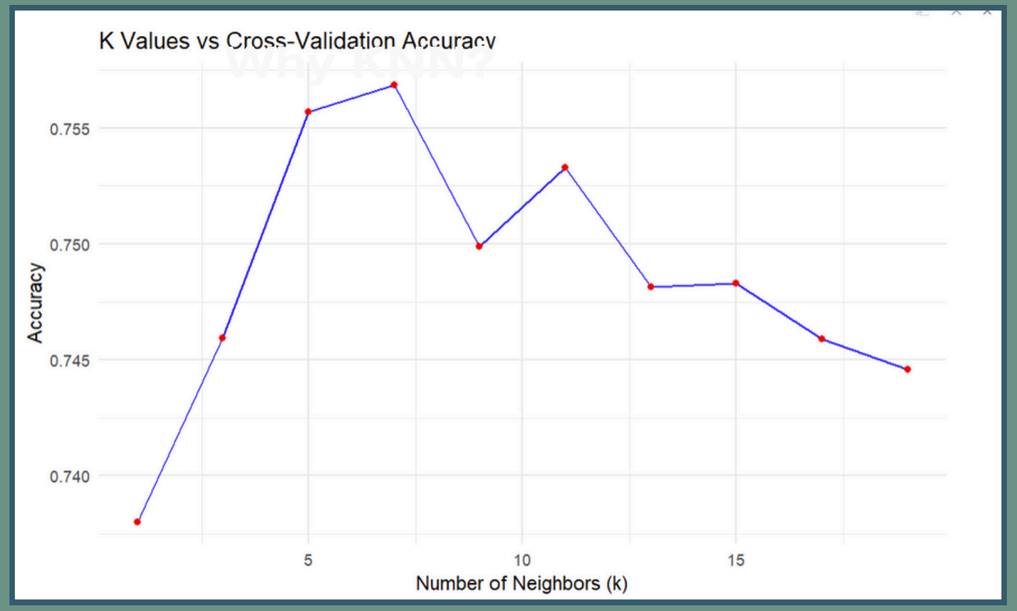
Top 5 Features	Top 5 Categories
Reviews	Finance
Price	Photography
Rating	Medical
Last Updated	Education
Content.Rating	House and Home

K-Nearest Neighbour



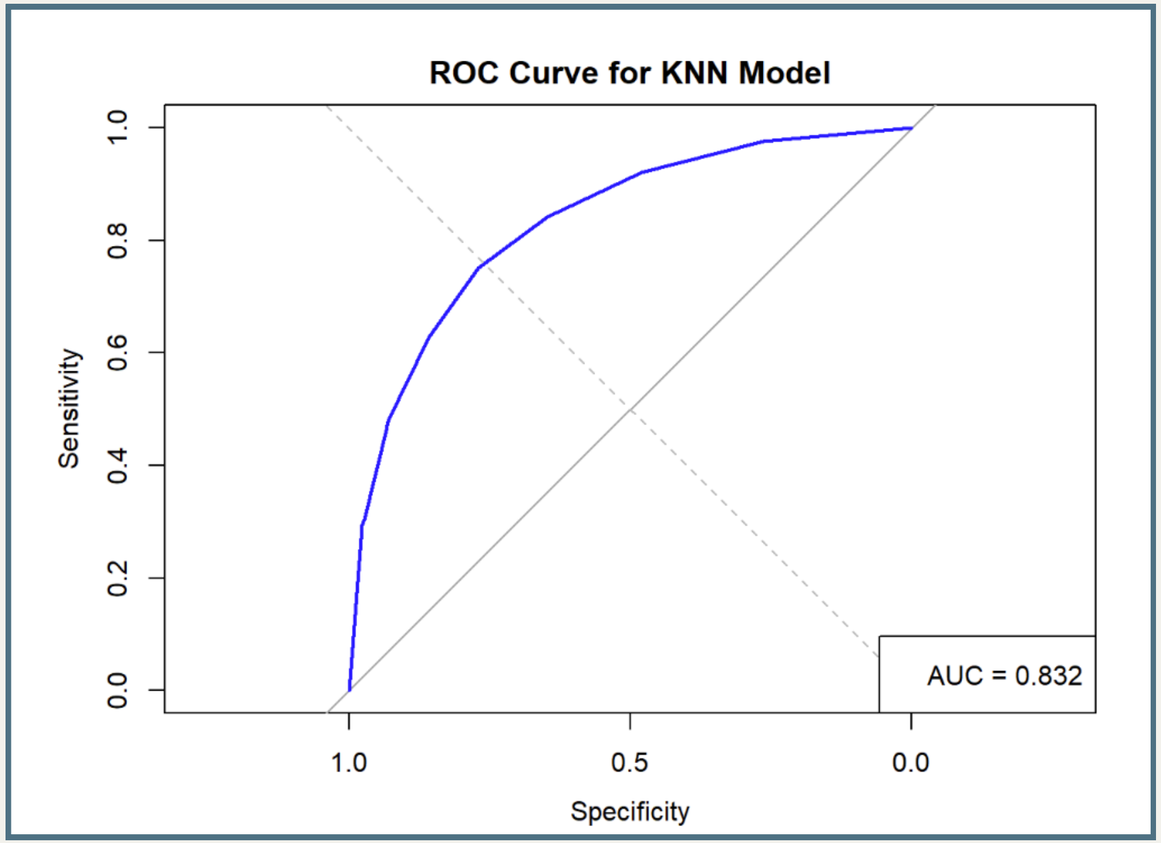
Model Development:

- Number of Neighbors (k): 7 (Optimal value determined through cross-validation)
- Number of Features: 39 (Scaled for better performance)
- Distance Metric: Euclidean distance



Model’s Performance Overview:

KNN Model ROC Curve:

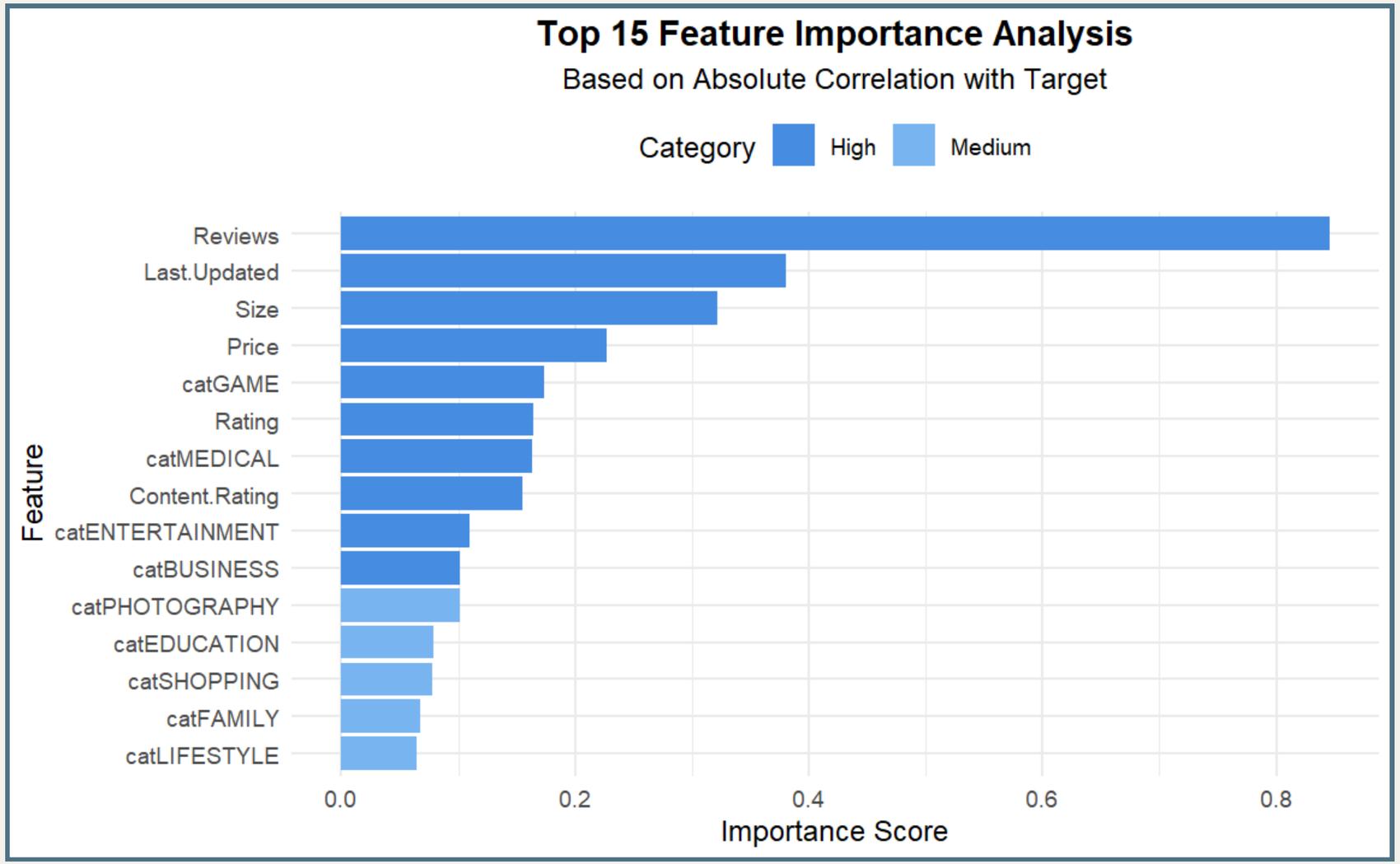


Actual

Confusion Matrix	0	1
Predicted 0	1361	371
Predicted 1	406	1114

Metric	Percentage
Accuracy	76.1
Precision	78.57
F1 - Score	77.79
Recall	77.02

Feature Importance:

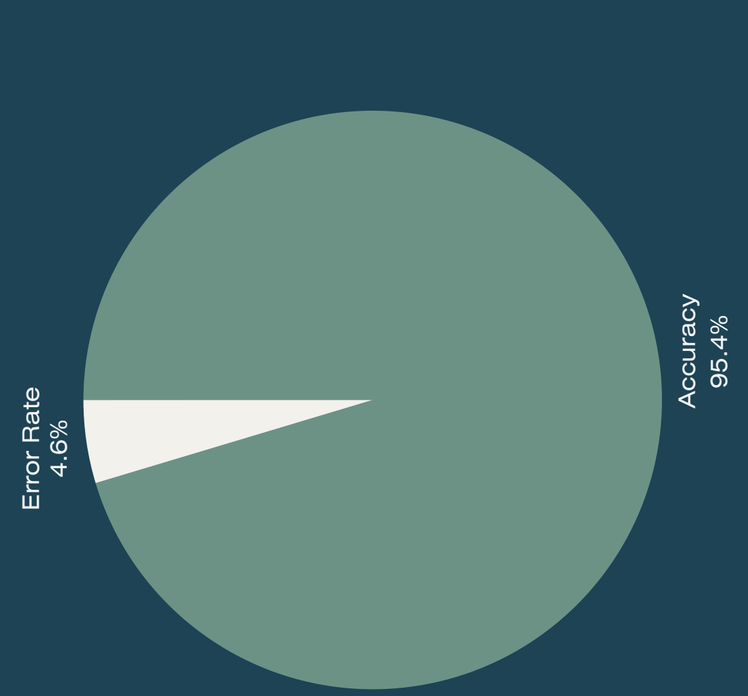


- **Highest Impact Features:**
  - Features like Reviews, Last Updated, and Size have the highest importance, indicating they strongly influence the classification of apps into low or high installs.
- **Category Contributions:**
  - Features derived from categorical variables (e.g., catGAME, catBUSINESS) show moderate importance, reflecting the varying impact of app categories.

Random Forest Classifier

Model Development

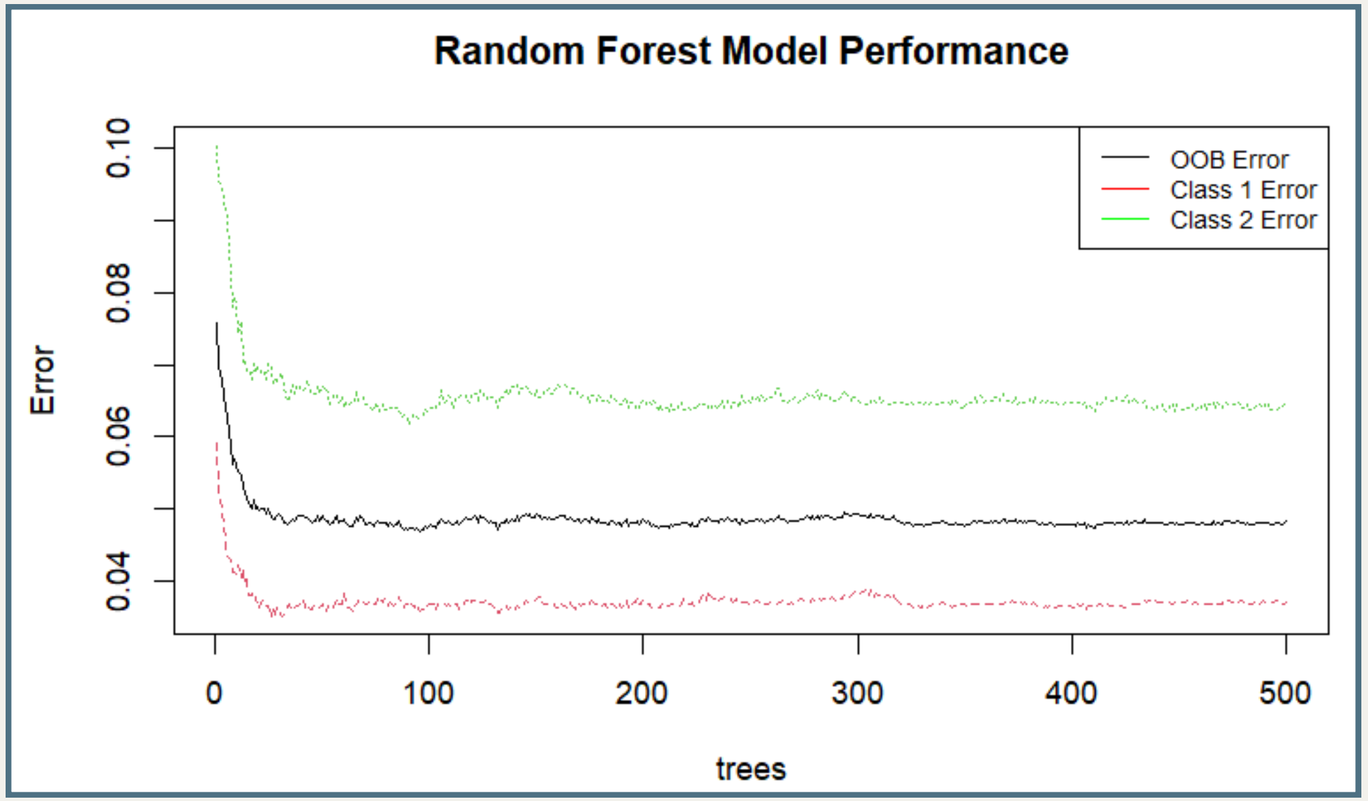
Number of Trees: 500
Number of Predictors: 39
(Excluding the target variable)
Variables per Split: 6 (Determined by mtry = sqrt(ncol(trainData) - 1))



Hyperparameter tuning:

	Accuracy	Kappa
mtry 2	93.53%	0.8647
mtry 4	95.4%	0.8995
mtry 6	95.17%	0.8995
mtry 8	95.08%	0.8976
mtry 10	95.10%	0.8981

Model’s Performance Overview:

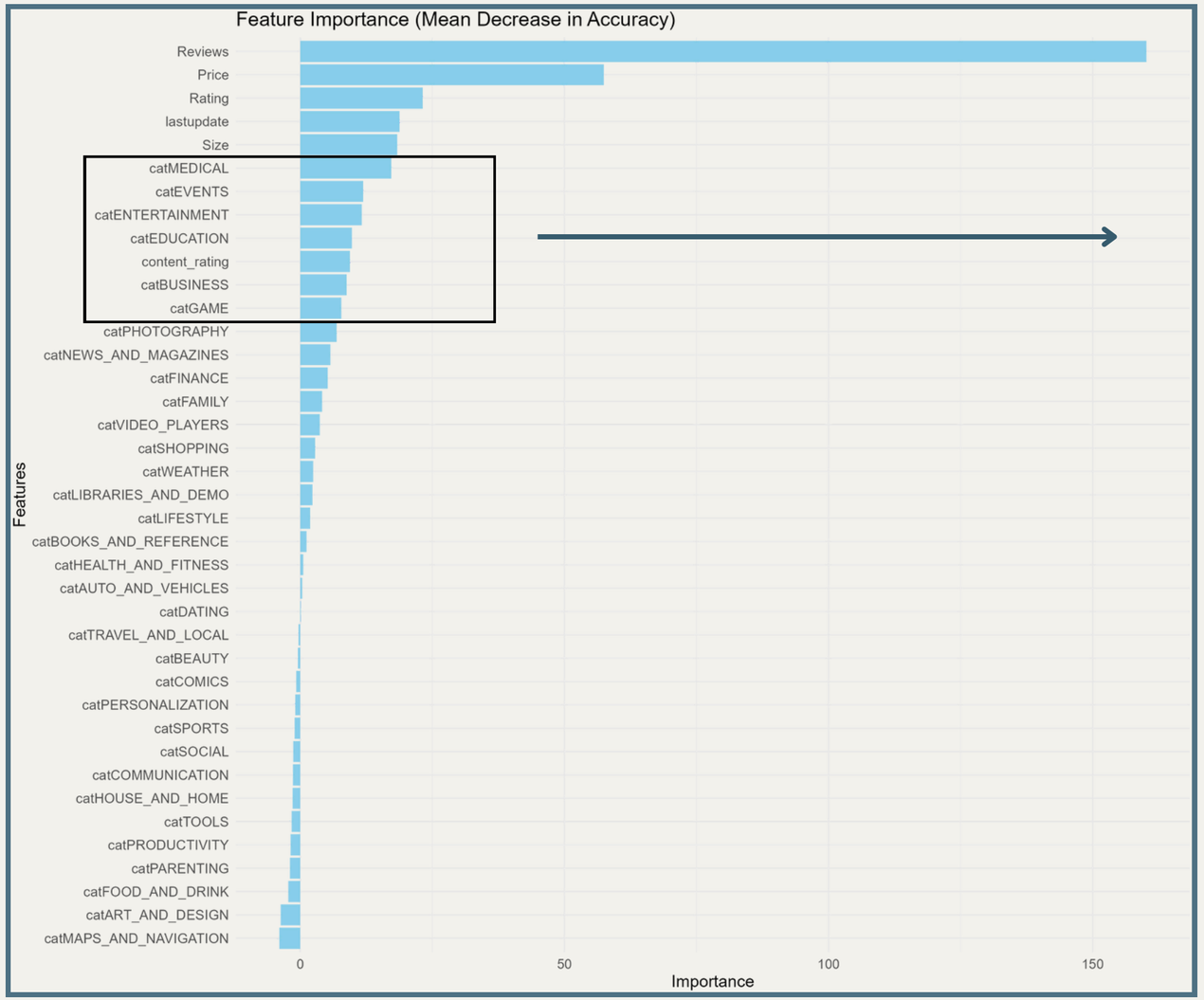


Actual

Confusion Matrix	0	1
Predicted 0	4425	171
Predicted 1	202	2918

DATASETS	ACCURACY
Training	0.963
Testing	0.954
OOB	0.952

Feature Importance Plot

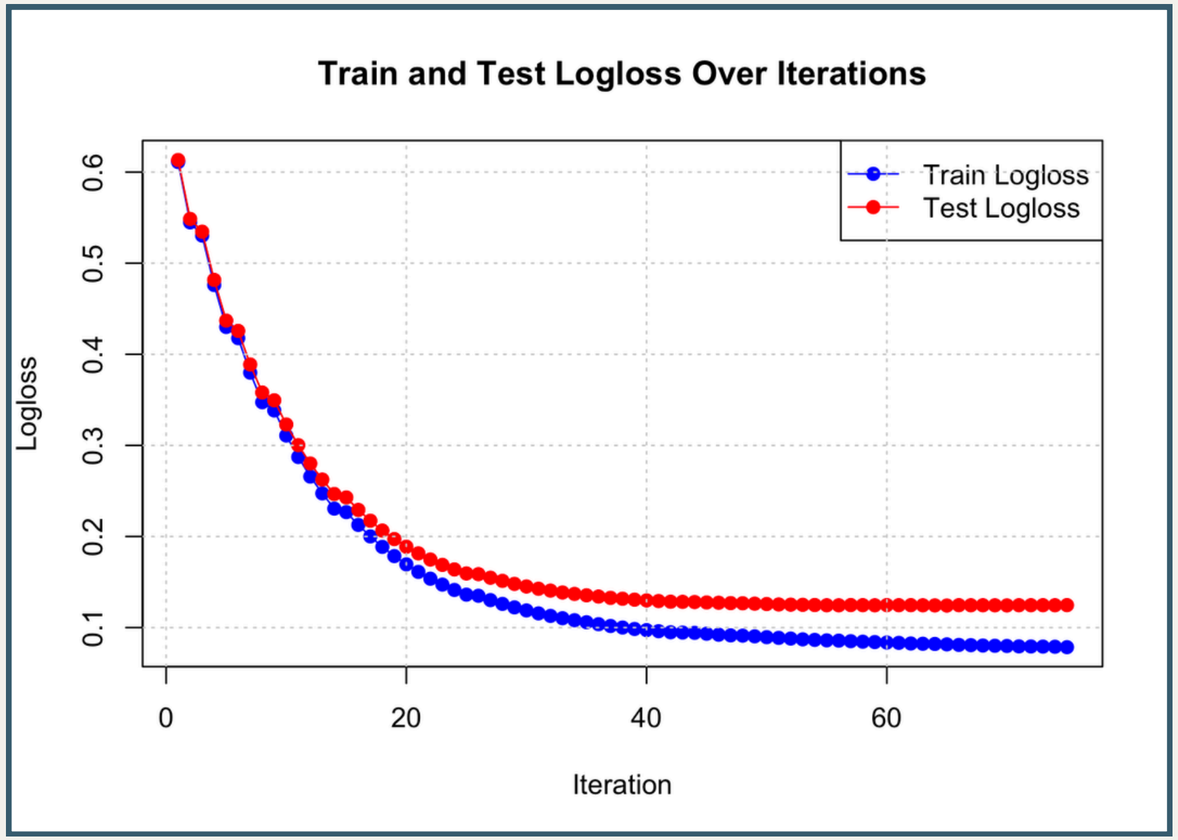


Gradient Boost

Why Gradient Boosting ?

Boosting is ideal for this classifier model due to its ability to handle complex relationships, feature importance insights, robustness to outliers, and high predictive power, ensuring accurate predictions and effective handling of mixed data types.

Evaluation of the model

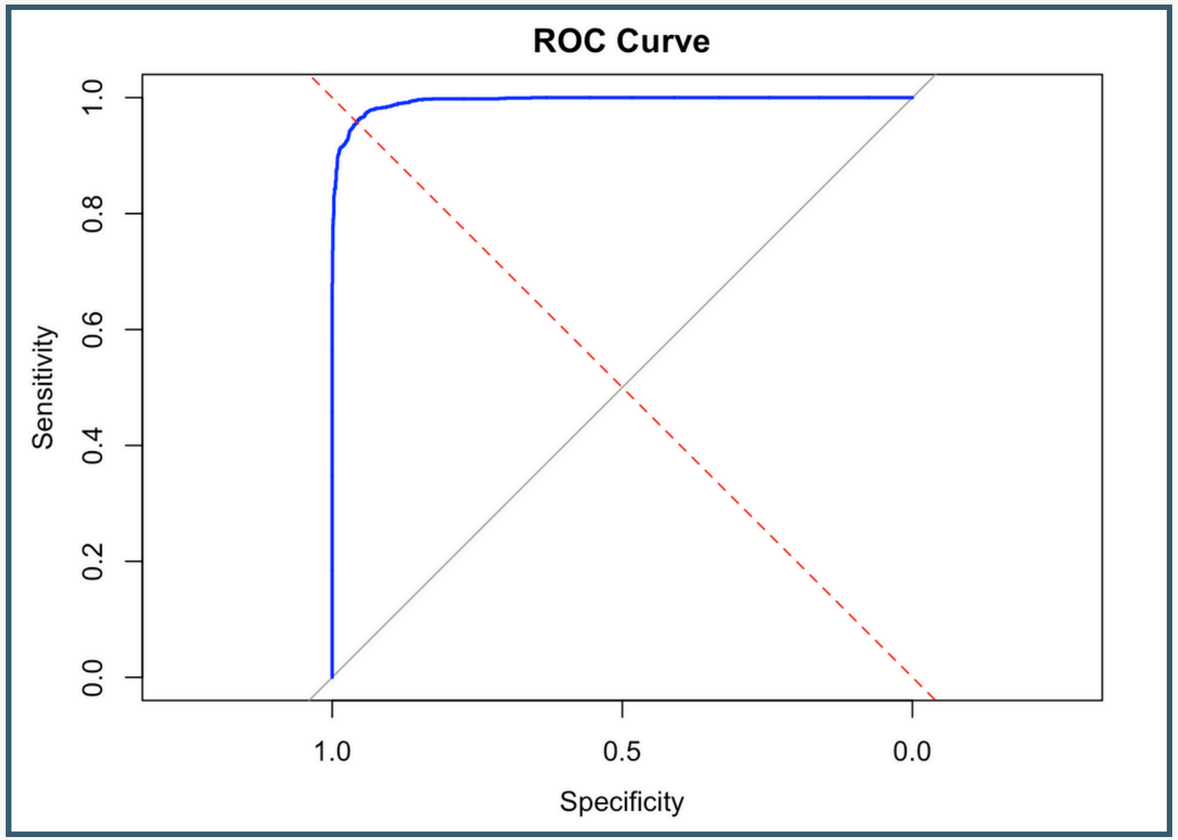


Parameters

objective	Binary Classification
Evaluation metric	Logloss
Max Depth	6
ETA(Learning Rate)	0.1
Sub sample	0.8
colsample_bytree	0.8

Metric	*Best Iteration*
Train Logloss	0.60
Test Logloss	0.61

Model’s Performance Overview:

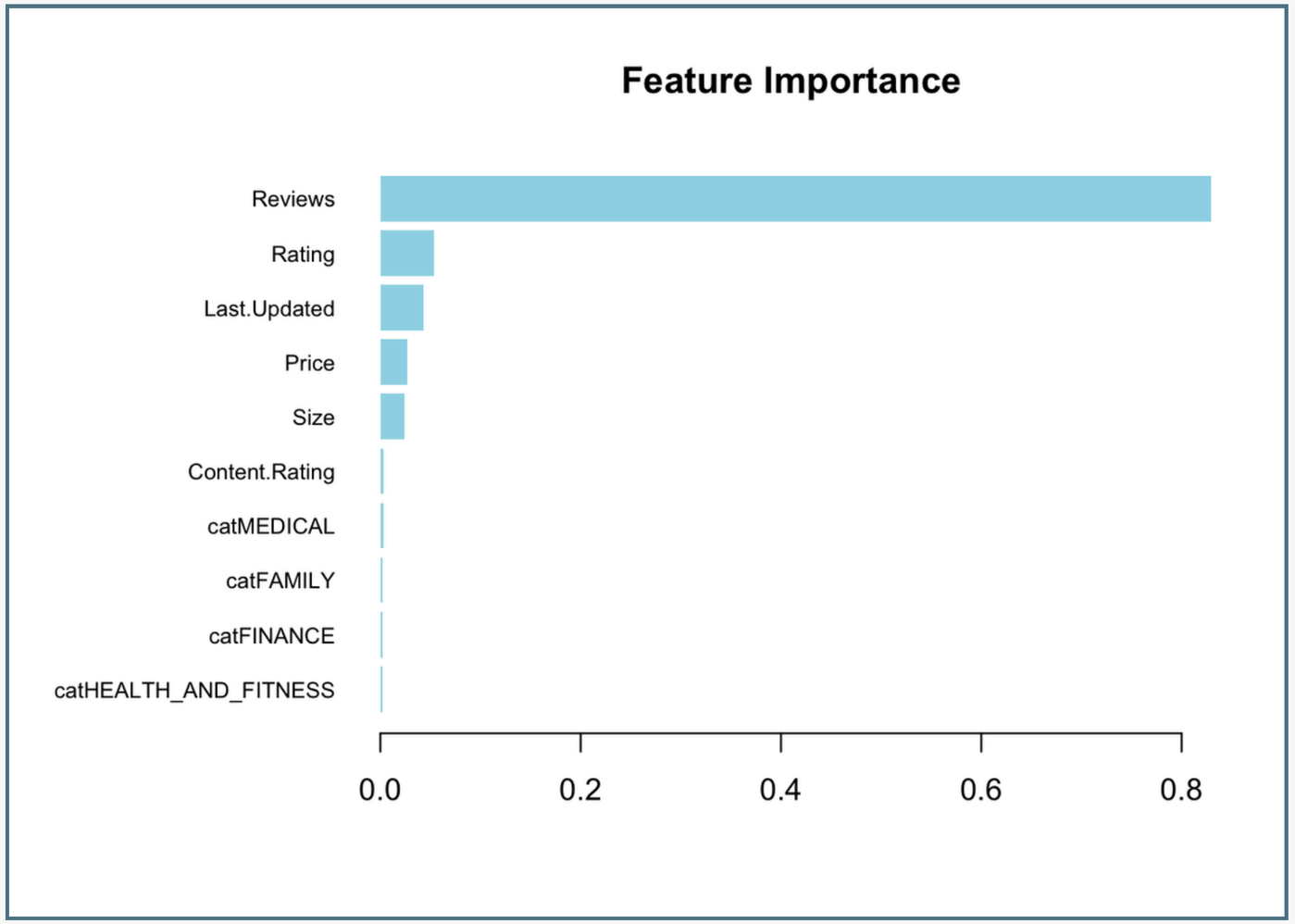


Actual

Confusion Matrix	0	1
Predicted 0	1702	71
Predicted 1	65	1414

Metric	Percentage
Accuracy	95
Precision	96

Feature Importance Plot



Top 5 features	**Top 5 Categories**
Reviews	Medical
Rating	Family
Last Updated	Finance
Price	Fitness
Size	News and Magazines

Support Vector Machine(Linear)

Higher Accuracy in Linear SVM :
The linear SVM model achieved greater accuracy than the non-linear model, indicating that the target variable is linearly separable.

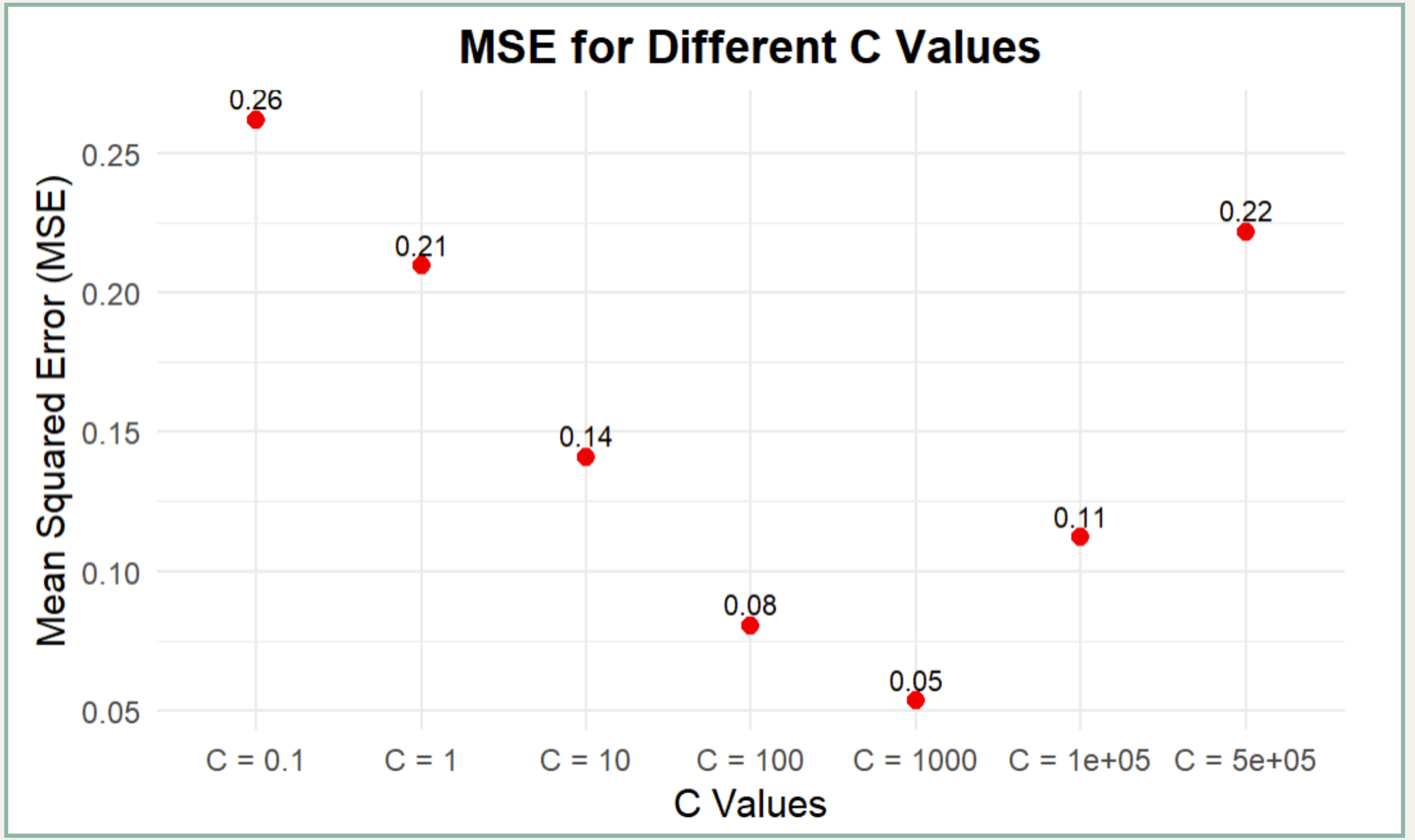
Model Development

C value : 1000
Number of Predictors: 39
(Excluding the target variable))

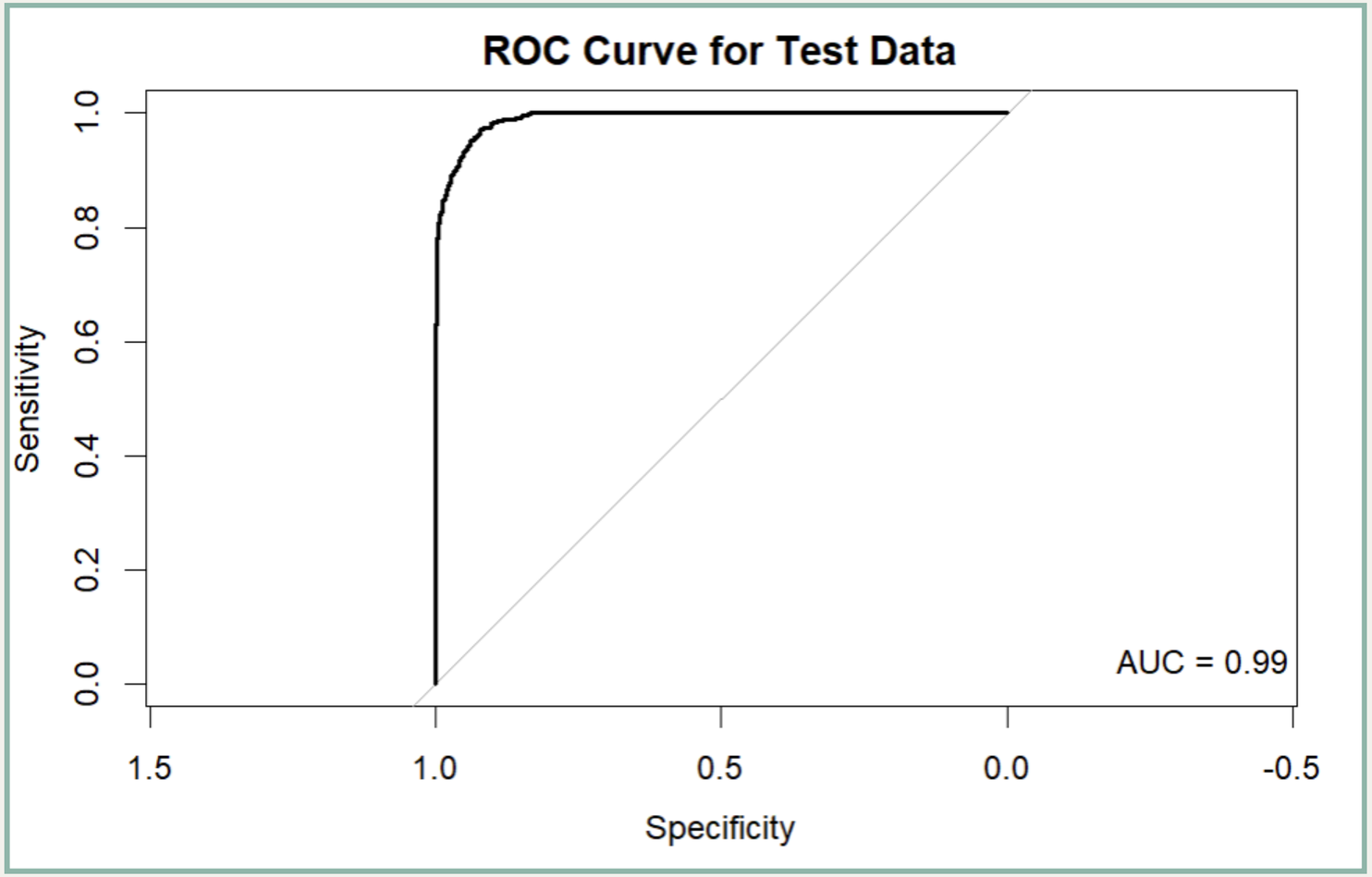
Hyperparameter tuning using K-fold Cross Validation:

C	Accuracy	Kappa
1	78.41%	0.554
10	84.63%	0.682
100	91.97%	0.815
1000	94.8%	0.895

Finding Optimal C for Test Data



Model’s Performance Overview:

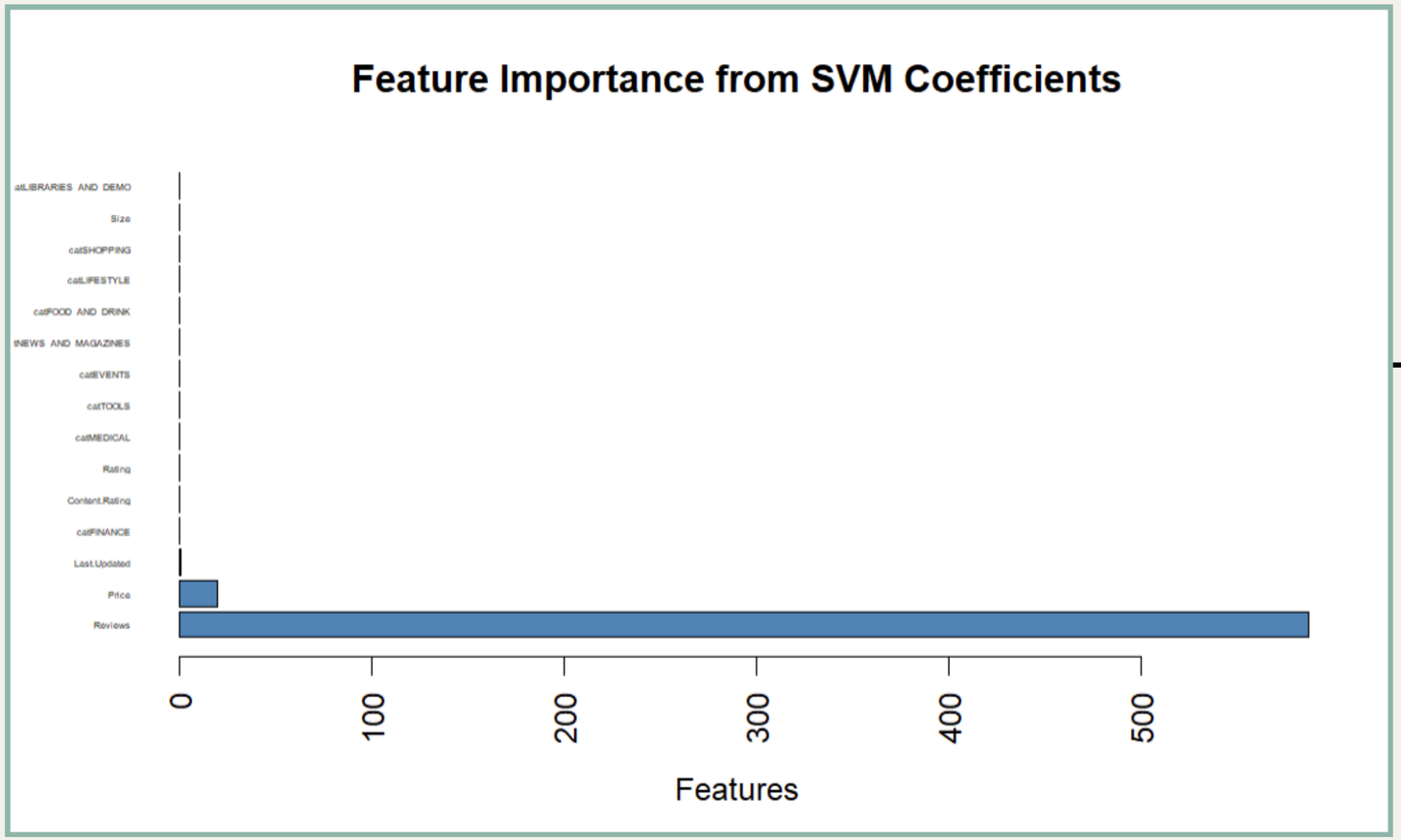


Actual

Predicted	Confusion Matrix	*0*	**1**
	0	1441	114
	1	31	1123

Metric	*Percentage*
Accuracy	94.64
Precision	92.66
F1 - Score	95.20
Recall	97.89

Feature Importance



Top 5 Categories	*Top 5 features*
Finance	Reviews
Medical	Price
Tools	Last Updated
Events	Finance
News and Magazines	Content Rating

Comparision between Model's Performance

METRIC/ ATTRIBUTE	LOGISTIC REGRESSION	KNN	RANDOM FOREST	GRADIENT BOOST	SVM
Accuracy	0.948	0.76	0.9517	**0.9581**	0.946
Precision	0.933	0.7857	0.955	**0.96**	0.926
Recall	0.973	0.7702	0.962	0.96	**0.978**
F1 Score	0.953	0.7779	0.959	**0.96**	0.952
AUC-ROC Score	**0.990**	0.832	0.981	**0.99**	**0.99**
Overfitting Tendency	Low	Low	Moderate	Moderate	Low
Interpretability	High	Moderate	Moderate	Moderate	Moderate

Conclusion

Developers and stakeholders can use these insights to prioritize features and categories that matter most for driving installs. For instance:

- Enhancing user reviews and ratings.
- Optimizing app size
- Prefers a low-pricing strategy.
- Developing apps in high-importance categories like GAMES, MEDICAL, ENTERTAINTMENT, FINANCE, PHOTOGRAPHY for better market traction.

Thank You