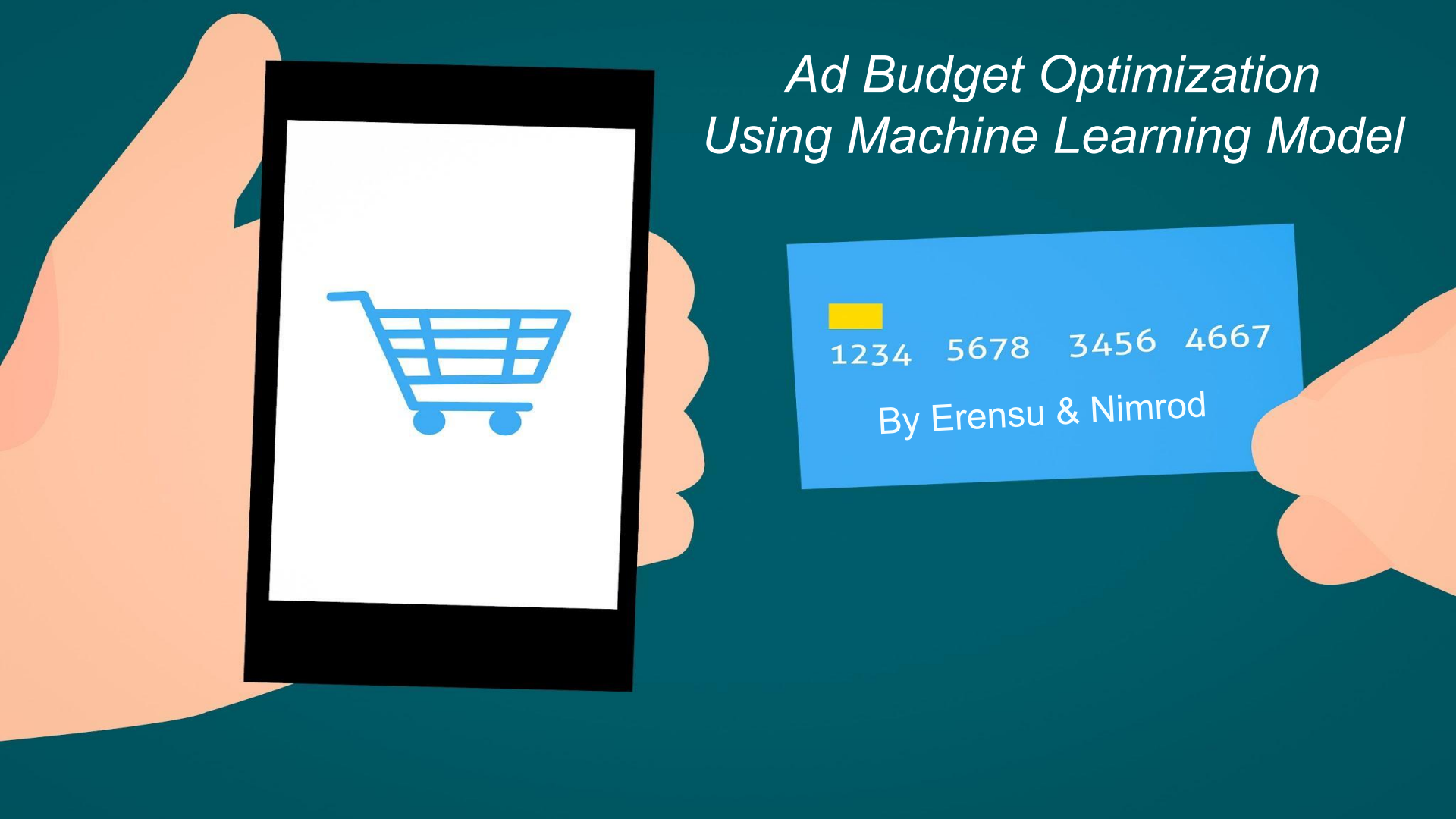
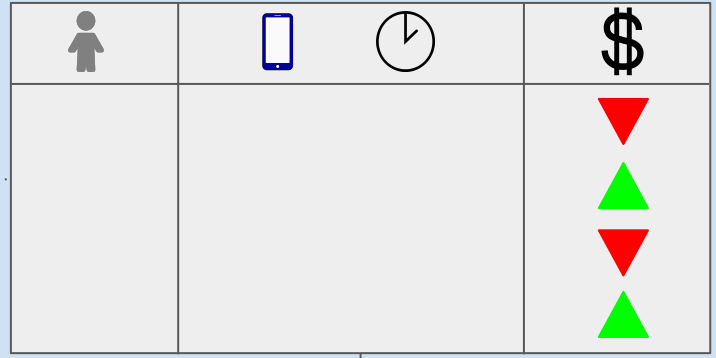
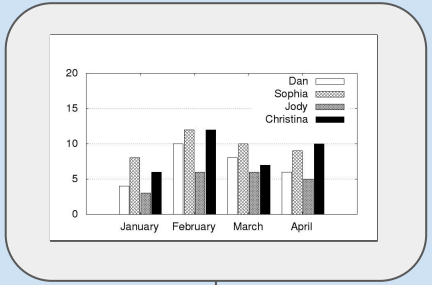
Ad Budget Optimization Using Machine Learning Model

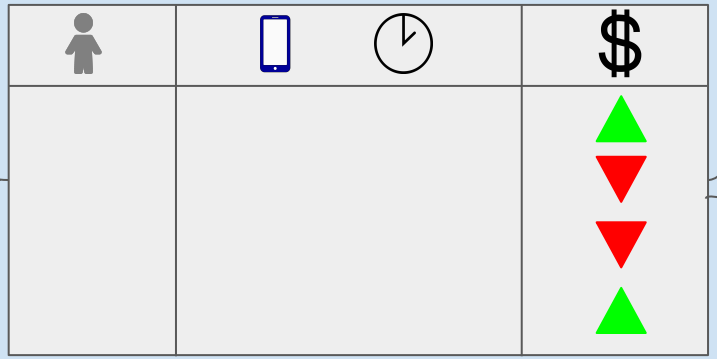


Older customers





New customers



Data Cleaning

Numerical

Derived new features by aggregating existing data and removed redundant attributes specifically within the numerical columns, enhancing the dataset for predictive modeling purposes.

- Added total value & order columns
- Monthly spending
- Order Frequency

Categorical

Applied one-hot encoding to convert categorical variables into numerical format

Used dictionary-based translation for categorical data transformation and get_dummies for encoding

- Used get_dummies in both cases
- Customer interest is represented as a list of interests, where more than one column may have a value of 1 to indicate multiple interests

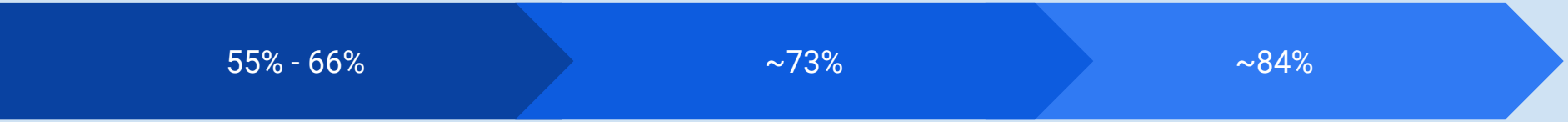
Dates

Applied pandas datetime functions to standardize date columns (e.g., first_order_date, last_order_date).

Addressed missing values and infinite values in order frequency per month, ensuring data integrity.

- Calculated total customer lifetime, time from last order, active days, and order frequency per month.
- Removed redundant attributes

Model Accuracy Process

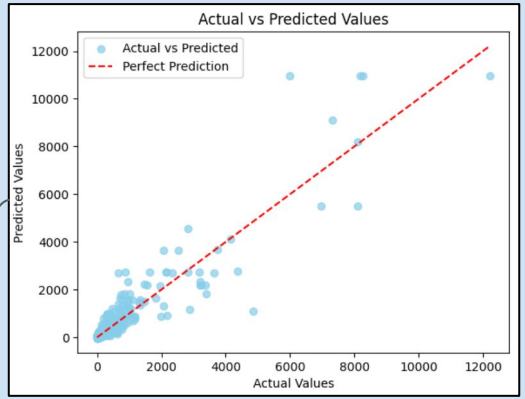


Achieved R^2 scores ranging from 0.55 to 0.66 by employing diverse scaling and normalization techniques on numerical variables, exploring various combinations for optimal model performance

Surprisingly managed to push accuracy further up by not scaling any variable with combination of using IQR for removal of outliers

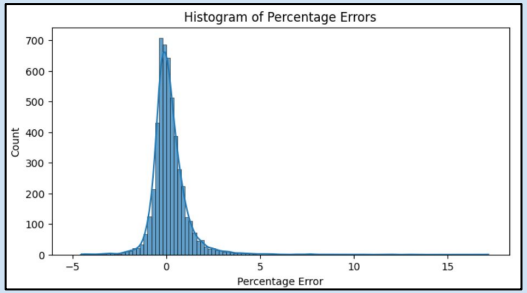
Finally, achieved 0.84 by converting the dependent variable to monthly profit from total profit. (total_value)

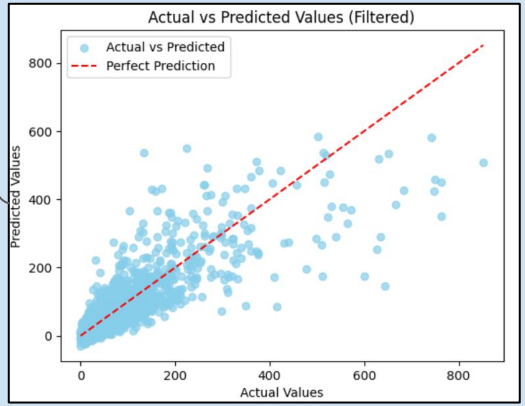
Model Accuracy Results

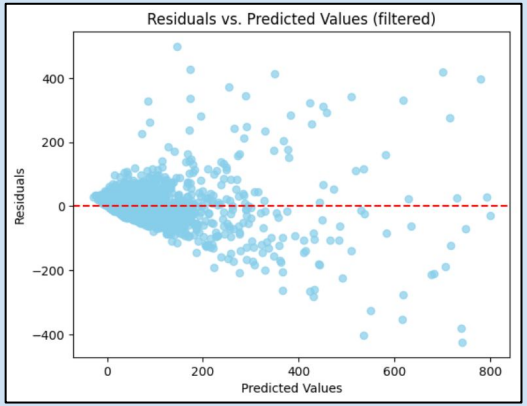


percentage_errors =

predicted_values - actual_values / actual_values







Conclusion & Issues

- The R-squared value is 0.842, indicating that approximately 84.2% of the variability in monthly profitability is explained by the independent variables in the model.
- @ 0.05 level, statistically significant variables are:
  - total_num_orders
  - active_days
  - order_freq_per_month
  - children
  - sports
- Multicollinearity: independent variables are highly correlated.
- To predict, at least one month of customer history is needed .

OLS Regression Results

Dep. Variable:	monthly_profitability	R-squared:	0.842
Model:	OLS	Adj. R-squared:	0.841
Method:	Least Squares	F-statistic:	6618.
Date:	Thu, 08 Feb 2024	Prob (F-statistic):	0.00
Time:	11:16:56	Log-Likelihood:	-1.0186e+05
No. Observations:	14958	AIC:	2.037e+05
Df Residuals:	14945	BIC:	2.038e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	17.6729	16.395	1.078	0.281	-14.463	49.809
total_num_orders	-2.4590	0.637	-3.859	0.000	-3.708	-1.210
total_lifetime_days	-0.0057	0.006	-0.930	0.352	-0.018	0.006
time_from_last_order_days	-0.0188	0.012	-1.596	0.110	-0.042	0.004
active_days	0.0132	0.006	2.059	0.040	0.001	0.026
order_freq_per_month	182.7324	0.658	277.740	0.000	181.443	184.022
children	-10.9892	4.668	-2.354	0.019	-20.140	-1.838
children_sports	-0.7985	5.003	-0.160	0.873	-10.604	9.007
men	-4.5654	3.927	-1.163	0.245	-12.262	3.132
sports	22.8410	3.725	6.131	0.000	15.539	30.143
women	-5.3287	3.770	-1.413	0.158	-12.719	2.062
android_app	5.3907	4.692	1.149	0.251	-3.806	14.588
desktop	8.9868	6.171	1.456	0.145	-3.109	21.082
ios_app	9.1158	5.765	1.581	0.114	-2.185	20.416
mobile	-5.8204	5.148	-1.131	0.258	-15.910	4.270

Omnibus:	27326.265	Durbin-Watson:	2.003
Prob(Omnibus):	0.000	Jarque-Bera (JB):	221808165.706
Skew:	12.766	Prob(JB):	0.00
Kurtosis:	599.018	Cond. No.	2.79e+19

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 1e-28. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
R-squared on testing set: 0.8416091424723489

THANK YOU