FINAL PROJECT MACHINE LEARNING: PREDICTING CUSTOMER REVENUE

BY AKITHA PASANDUL

CONTENT

- Introduction
- About the dataset
- Machine learning analysis
- Key Findings and insights
- Model flaws

INTRODUCTION

- This project aims to predict customer lifetime value using linear regression on a Telco dataset. Key attributes like demographics, service usage and account information will be explored. After cleaning and feature engineering, simple, polynomial and regularized regressions will be trained and evaluated, focusing on accuracy and interpretability to identify factors driving total revenue.
- Insights will guide strategies to increase customer tenure and optimize service bundles, informing targeted marketing. Future steps involve exploring advanced models, interaction effects and the impact of churn, enhanced by additional data to improve prediction and provide actionable recommendations for the telecommunications company.

ABOUT THE DATASET

The Telco Customer Churn dataset contains rich information about telecommunication customers. Each row represents a unique customer and includes demographic details such as gender, age, and location, service information like phone and internet plans, and account details encompassing tenure, contract type, charges, and revenue. Customer experience metrics like satisfaction score, churn status, customer lifetime value (CLTV),and detailed churn reasons are also provided.

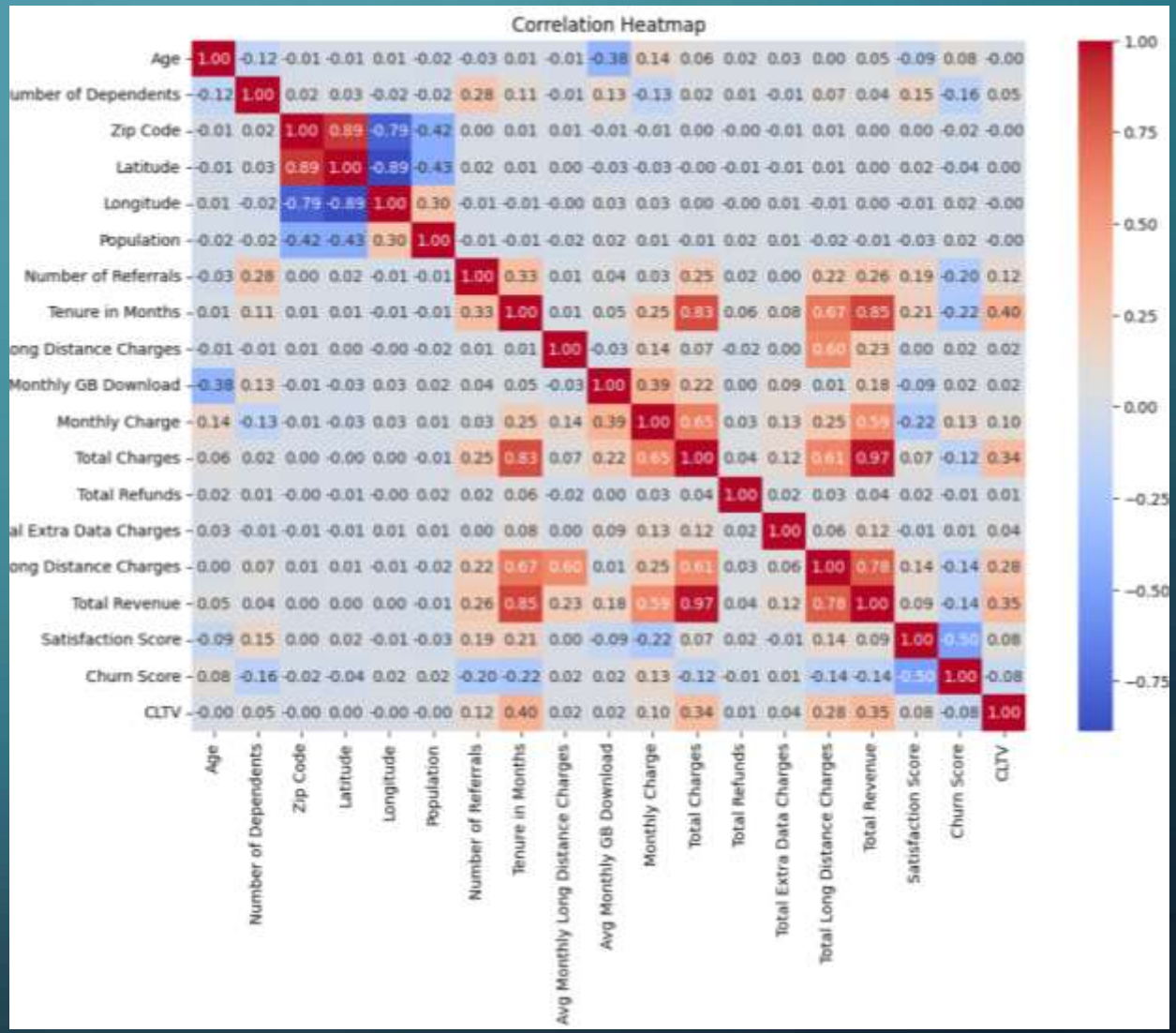
	Customer ID	Gender	Age	Under 30	Senior Citizen	Married	Dependents	Number of Dependents	Country	State	...	Total Extra Data Charges	Total Long Distance Charges	Total Revenue	Satisfaction Score	Customer Status	Churn Label	Churn Score	CLTV	Churn Category	Churn Reason
0	8779-QRDMV	Male	78	No	Yes	No	No	0	United States	California	...	20	0.00	59.65	3	Churned	Yes	91	5433	Competitor	Competitor offered more data
1	7495-OOKFY	Female	74	No	Yes	Yes	Yes	1	United States	California	...	0	390.80	1024.10	3	Churned	Yes	69	5302	Competitor	Competitor made better offer
2	1658-BYGOY	Male	71	No	Yes	No	Yes	3	United States	California	...	0	203.94	1910.88	2	Churned	Yes	81	3179	Competitor	Competitor made better offer
3	4598-XLKNJ	Female	78	No	Yes	Yes	Yes	1	United States	California	...	0	494.00	2995.07	2	Churned	Yes	88	5337	Dissatisfaction	Limited range of services
4	4846-WHAFZ	Female	80	No	Yes	Yes	Yes	1	United States	California	...	0	234.21	3102.36	2	Churned	Yes	67	2793	Price	Extra data charges

5 rows × 50 columns

ABOUT THE DATASET

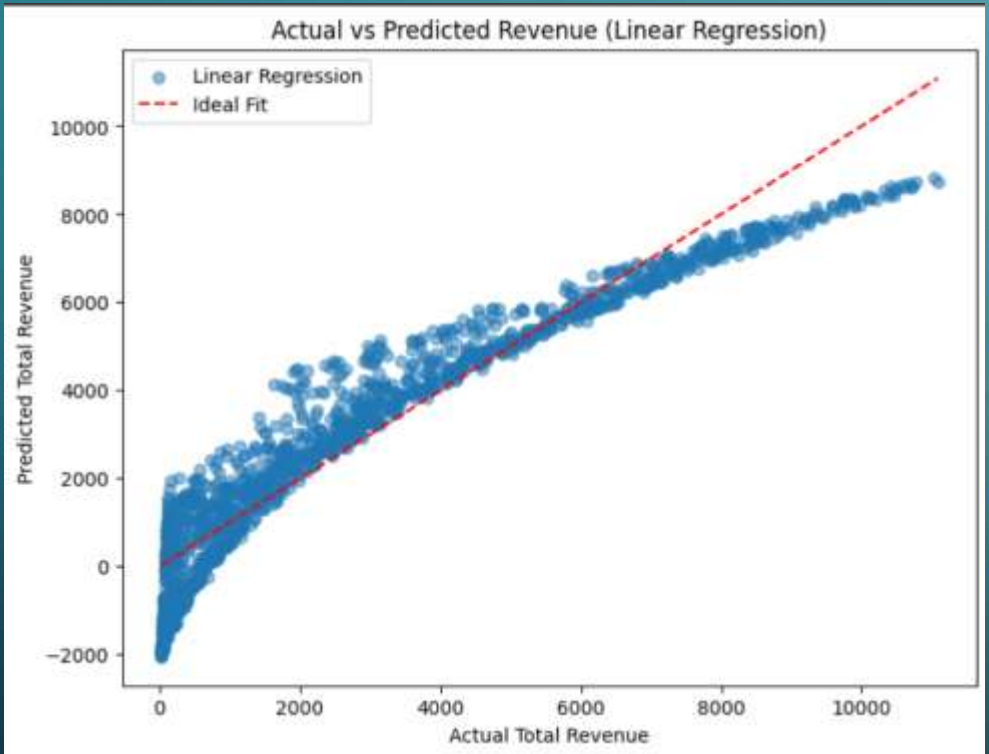
	Age	Number of Dependents	Zip Code	Latitude	Longitude	Population	Number of Referrals	Tenure in Months	Avg Monthly Long Distance Charges	Avg Monthly GB Download	Monthly Charge	Total Charges	Total Refunds	Total Extra Data Charges	Total Long Distance Charges	Total Revenue	Satisfaction Score	Churn Score	CLTV
count	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00	7043.00
mean	46.51	0.47	93486.07	36.20	-119.76	22139.60	1.95	32.39	22.96	20.52	64.76	2280.38	1.96	6.86	749.10	3034.38	3.24	58.51	4400.30
std	16.75	0.96	1856.77	2.47	2.15	21152.39	3.00	24.54	15.45	20.42	30.09	2266.22	7.90	25.10	846.66	2865.20	1.20	21.17	1183.06
min	19.00	0.00	90001.00	32.56	-124.30	11.00	0.00	1.00	0.00	0.00	18.25	18.80	0.00	0.00	0.00	21.36	1.00	5.00	2003.00
25%	32.00	0.00	92101.00	33.99	-121.79	2344.00	0.00	9.00	9.21	3.00	35.50	400.15	0.00	0.00	70.55	605.61	3.00	40.00	3469.00
50%	46.00	0.00	93518.00	36.21	-119.60	17554.00	0.00	29.00	22.89	17.00	70.35	1394.55	0.00	0.00	401.44	2108.64	3.00	61.00	4527.00
75%	60.00	0.00	95329.00	38.16	-117.97	36125.00	3.00	55.00	36.39	27.00	89.85	3786.60	0.00	0.00	1191.10	4801.15	4.00	75.50	5380.50
max	80.00	9.00	96150.00	41.96	-114.19	105285.00	11.00	72.00	49.99	85.00	118.75	8684.80	49.79	150.00	3564.72	11979.34	5.00	96.00	6500.00

CORRELATION BETWEEN THE FEATURES



MACHINE LEARNING ANALYSIS

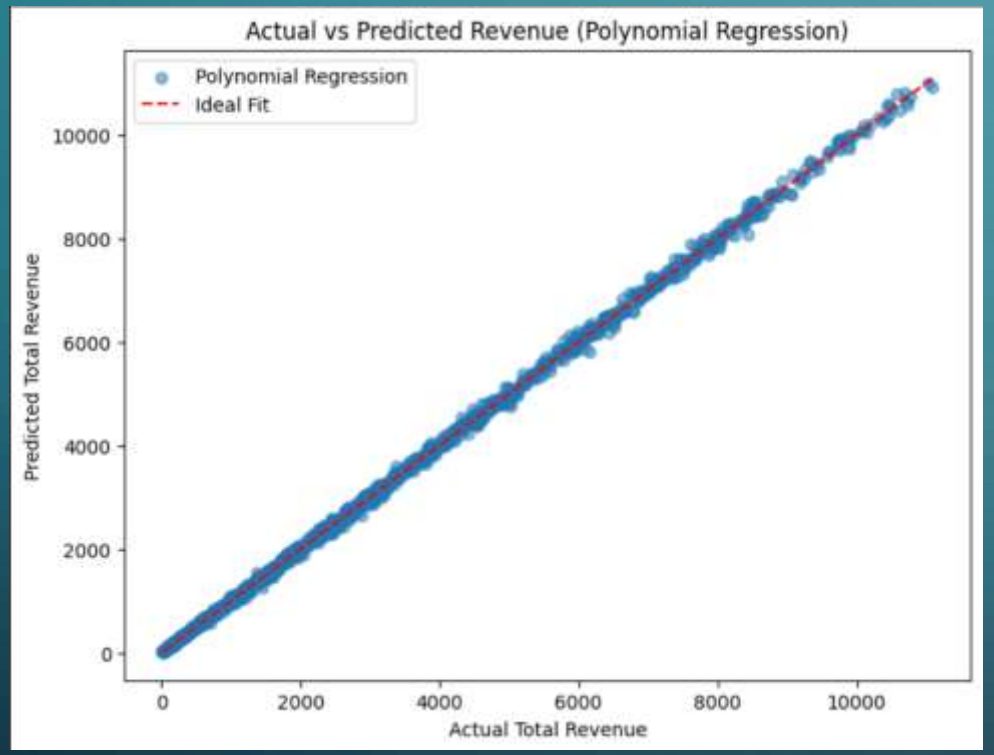
SIMPLE LINEAR REGRESSION



- MSE: 768451.7894
- R2: 0.9034

- Have a relatively higher MSE (Mean Squared Error) and a lower R-squared (R^2) compared to the other models, indicating lower accuracy in predicting Total Revenue.

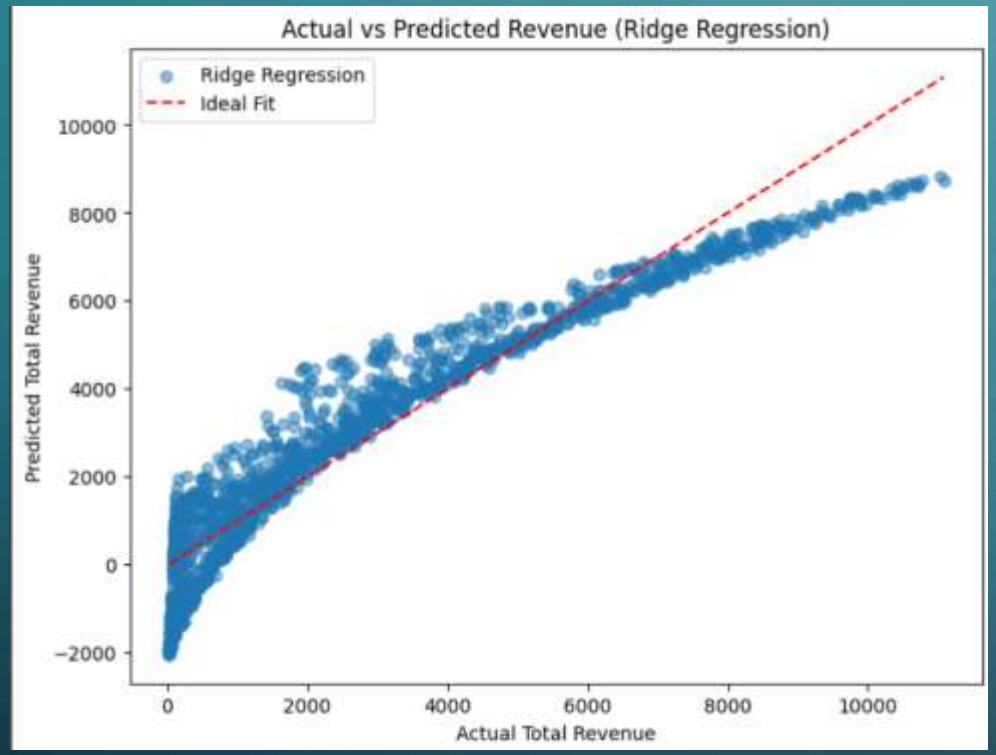
POLYNOMIAL REGRESSION



- MSE: 4821.8395
- R2: 0.9993

- A lower MSE and a higher R-squared compared to the simple linear regression, suggesting improved prediction accuracy.
- However, there is a risk of overfitting especially if the polynomial degree is too high.

RIDGE REGRESSION



- MSE: 768402.5709
- R2: 0.9039

- The MSE and R-squared are expected to be better than simple linear regression and potentially comparable to or slightly better than polynomial regression

MODEL COMPARISON

•The best model will be chosen based on a balance between prediction accuracy and interpretability. Ridge regression often strikes a good balance, providing reasonably accurate predictions with more stable and interpretable coefficients.

•If the polynomial regression significantly outperforms Ridge, it might be worth considering despite the complexity, but careful attention should be paid to potential overfitting.

	Model	MSE	R2
0	Linear Regression	768451.79	0.90
1	Polynomial Regression	4821.84	1.00
2	Ridge Regression	768402.57	0.90

KEY FINDINGS AND INSIGHTS

KEY FINDINGS

- **Tenure is a Dominant Driver:** Customer tenure emerges as a primary factor influencing total revenue. Longer-tenured customers contribute significantly more to overall revenue. This suggests that customer retention strategies are crucial for maximizing long-term profitability.
- **Service Bundles Matter:** Customers subscribing to comprehensive service bundles tend to generate higher total revenue. Offering attractive and value-added service bundles can encourage customers to adopt more services, leading to increased revenue.
- **Monthly Charge as an Indicator:** Higher monthly charges are positively correlated with total revenue, but not always linearly. While premium services increase revenue, excessive charges might deter customers or lead to churn. Balancing pricing and perceived value is essential.
- **Impact of Demographics:** Senior citizens, although potentially on fixed incomes, may represent a valuable customer segment due to their higher adoption of specific services or longer tenure. Tailoring services and marketing strategies to this demographic could be beneficial.
- **Location-Based Insights:** Geographic location can influence revenue, with certain areas exhibiting higher spending patterns. Understanding regional preferences and tailoring services accordingly can optimize revenue generation.
- **Offers and Contractual Agreements:** Specific promotional offers or contract types may attract customers but could impact long-term revenue. Analyzing the performance of different offers and contract terms in relation to total revenue is important for optimizing promotional strategies.
- **Influence of Long-Distance Charges:** Total Long-Distance Charges might contribute to the total revenue. Customers engaging with long-distance calls might generate higher revenue.

MODEL FLAWS AND SOLUTIONS

SIMPLE LINEAR REGRESSION

Flaws:

- Oversimplification: Assumes a linear relationship between predictors and Total Revenue.
- potentially important variables and complex relationships.
- Sensitivity to Outliers: Outliers can disproportionately influence the regression line.

Solutions:

- Feature Engineering: Introduce new features that capture non-linear aspects, like squared terms, interaction terms or categorical variable combinations.
- Outlier Handling: Identify and mitigate the impact of outliers through techniques like minorizing or robust regression.
- Variable Selection: Carefully select the most relevant features based on domain knowledge and exploratory data analysis.
- Transformation: Try transforming the target variable using log transformation to reduce skewness.

POLYNOMIAL REGRESSION

Flaws:

- Overfitting: Can easily overfit the training data, leading to poor generalization on the test set. High-degree polynomials can create overly complex curves that fit noise.
- Multicollinearity: Polynomial terms introduce multicollinearity making coefficient interpretation difficult.
- Instability: Small changes in the data can lead to large changes in the fitted polynomial.

Solutions:

- Regularization: Combine polynomial regression with regularization techniques like Ridge or Lasso to penalize overly complex models and prevent overfitting.
- Cross-Validation: Use cross-validation to tune the degree of the polynomial and select the model with the best generalization performance.
- Feature Scaling: Scaling the original features before adding polynomial terms is crucial to mitigate multicollinearity.

RIDGE REGRESSION

Flaws:

- Feature Selection Limitation: Ridge regression shrinks coefficients but doesn't force them to be exactly zero. Therefore, it doesn't perform feature selection. It includes all predictors in the model, although with reduced influence.
- Model Complexity: Can be more difficult to interpret than simple linear regression, especially with many features.
- Assumption of Linearity: Still assumes a linear relationship after the regularization.

Solutions:

- Feature Engineering: Create polynomial or interaction terms.
- Variable Selection: Select the most relevant features based on domain knowledge and exploratory data analysis.
- Other Regularization Techniques: Combine it with other feature engineering techniques such as polynomial features.

THANK YOU