Supervised Machine Learning: Regression
Peer-Reviewed Course Project
M. Weston

I. Main Objective

The main objective of this project is to develop a predictive model for churn, measured by how many months the customer was with the credit card company. This project will use a free dataset of credit card customer churn. This dataset and predictive model goal were chosen for their similarity to a business project faced by the author, who is trying to estimate research panel tenure of voluntary research participants in media marketing measurement. The key features used in this dataset, like age and gender, are also important factors in the media marketing problem.

II. Data

This project will use the Credit Card Customers free dataset, which can be downloaded here:

Goyal, S. (2020, November 19). Credit Card Customers. Retrieved December 15, 2020, from https://www.kaggle.com/sakshigoyal7/credit-card-customers?select=BankChurners.csv

In this dataset, each row represents a different credit card customer. There are a total of 10,127 customers in this dataset. The columns are described in the table below. Some columns were removed from the file prior to modeling to reduce complication. A value of N/A is used when the list of possible values is too long for inclusion or the units are unknown or irrelevant. No readme was provided, so descriptions are assumed from column headers.

Column Name	Description	Possible Values	Units
CLIENTNUM	Customer ID	N/A	N/A
Attrition_Flag	If the customer is current or churn	Existing Customer, Attrited Customer	N/A
Customer_Age	Customer age	26-73	Years
Gender	Customer sex	M, F	N/A
Dependent_count	Number of dependents	0-5	Persons
Education_Level	Highest level of degree completed by customer	College, Doctorate, Graduate, High School, Post-graduate, Uneducated, Unknown	N/A

Marital Status	Marital status of customer	Divorced, Single, Married, Unknown	N/A
Income_Category	Household income of the customer	Less than $40K, $40K - $60K, $60K - $80K, $80K – 120K, $120K +, Unknown	U.S. dollars
Card_Category	Credit card category of the card holder	Blue, Gold, Platinum, Silver	N/A
Months_on_book	Tenure with the company	N/A	Months
Months_Inactive_12_mon	How many months in the last 12 the credit card was inactive	N/A	Months
Contacts_Count_12_mon	How many times in the last 12 months the customer contacted the credit company	N/A	N/A
Credit_Limit	Card credit limit	N/A	N/A
Total_Revolving_Bal	Total revolving balance for the credit card	N/A	N/A
Total_Trans_Amt	Total transaction amount credit card was used for	N/A	U.S. dollars
Total_Trans_Ct	Total number of transactions credit card was used for	N/A	N/A
Avg_Utilization_Ratio	Average utilization ratio for the credit card	N/A	Decimal percent

III. Data Exploration and Cleaning

Data Cleaning:
This dataset was cleaned by the creator. However, to ensure it was properly cleaned, the following items were confirmed:

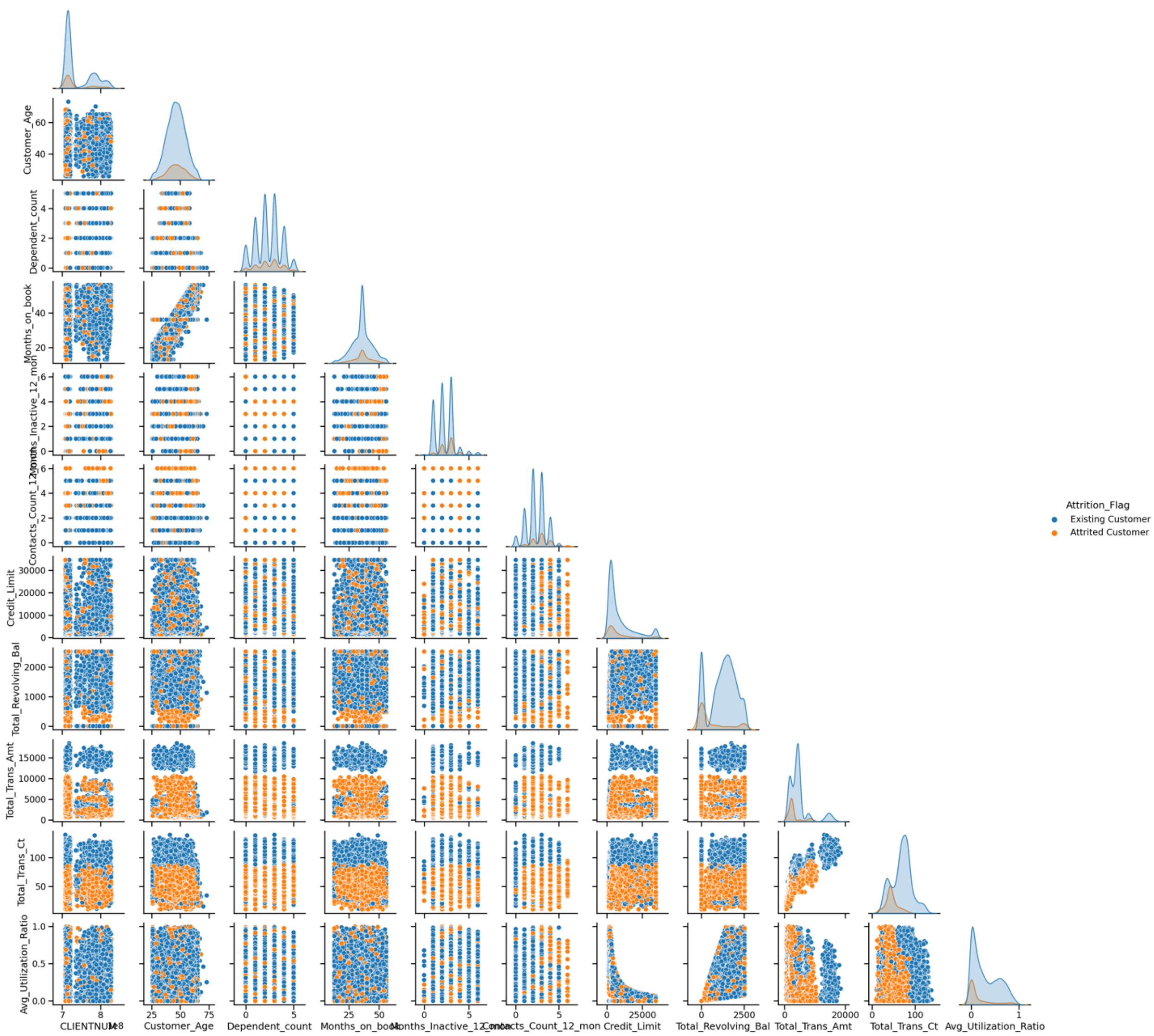
1. No customer IDs were repeated (no duplicate entries)
2. There are no missing values
3. Data types and formats are consistent across columns
4. No outliers were visually identified (see Data Exploration)

Data Transformations:
There are six columns that consist of string data and must be encoded. These changes are explained in the table below.

Variable	Possible Values	Encoding Method	Definitions
Attrition_Flag	Existing Customer, Attrited Customer	Binary	0 – Existing Customer 1 – Attrited Customer
Gender	M, F	Binary	0 – M 1 – F
Education_Level	College, Doctorate, Graduate, High School, Post-Graduate, Uneducated, Unknown	Ordinal	0 – Unknown 1 – Uneducated 2 – High School 3 – College 4 – Graduate 5 – Doctorate 6 – Post-Graduate
Marital_Status	Divorced, Single, Married, Unknown	One-hot encoding	No column created for unknown (null)
Income_Category	Less than $40K, $40K - $60K, $60K - $80K, $80K - $120K, $120K +, Unknown	Ordinal	0 – Unknown 1 – Less than $40K 2 – $40K - $60K 3 – $60K - $80K 4 – $80K – $120K 5 – $120K +
Card_Category	Blue, Gold, Platinum, Silver	Ordinal	0 – Blue 1 – Silver 2 – Gold 3 – Platinum

Data Exploration:
Data exploration was performed with Seaborn because it allows the user to compare multiple features simultaneously. The hue was set using the Attrition_Flag column to quickly identify if any features relate directly to churn. The Total_Trans_Ct feature seems most likely to correlate with churn, with attrited customers having lower transaction counts on average.

Note: This analysis was performed both before and after the column transformations. Both runs resulted in similar information in terms of relationships. The pre-transformation plot is given here for simplicity and ease of reading.



IV. Modeling

For all models, a regular training-testing split was used to save time. The split was defined as follows:

```
from sklearn.model_selection import train_test_split
train, test = train_test_split(data, test_size = 0.3)
```

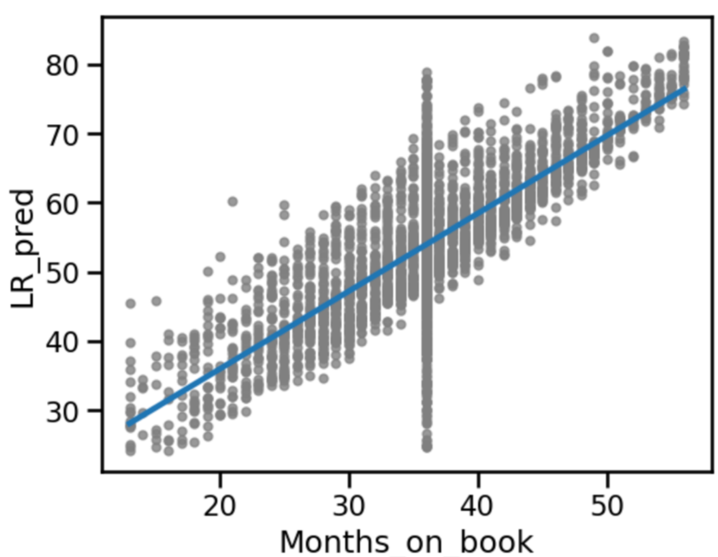
As mentioned above, the main objective of this project is to develop a predictive model for churn, measured by how many months the customer was with the credit card company. This will correspond to the Months_on_book column.

Simple Linear Regression Model

Linear regression assumes a normal distribution of the y-variable (in this case, Months_on_book). Three different methods of normalization were attempted for this variable. None of the methods returned a good normalization p-value, as seen in the table below, indicating that simple linear regression is unlikely to be an accurate model for this data set. However, linear modeling is completed using the best method, a Box-Cox transformation.

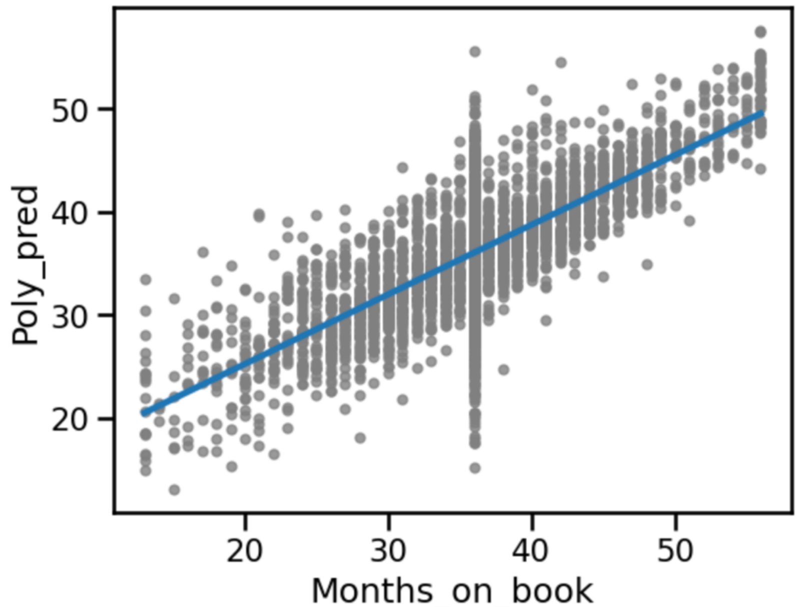
Normalization Method	Resulting p-value
No normalization	2.934e-15
Log normalization	0.0
Square root transformation	4.348e-141
Box-Cox transformation	5.972e-9

The comparison between the predicted and actual months of tenure is given below. As expected from the non-normalcy, the fit is horrible.



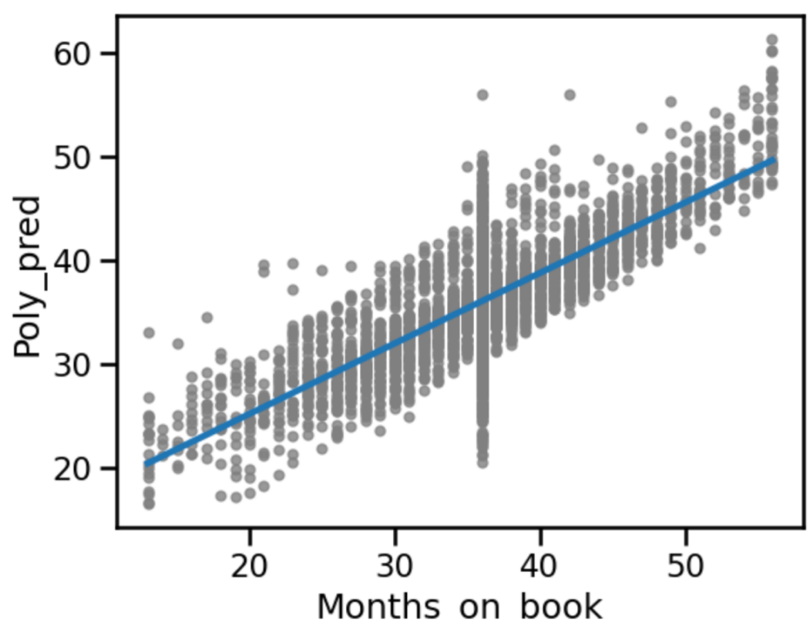
Second-Degree Polynomial Regression Model

A second-degree polynomial regression model returned a far better prediction.



Fourth-Degree Polynomial Regression Model

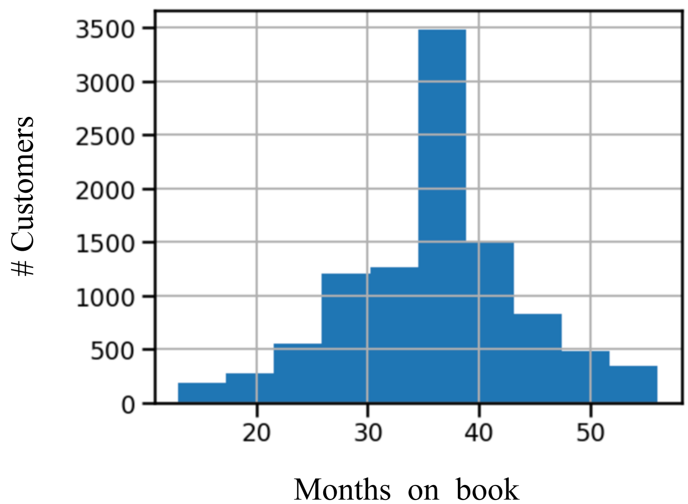
Because the second-degree polynomial regression model produced such promising results, a higher-order polynomial regression model is used for the third model evaluated in this analysis. It showed marginal improvement over the second-degree model.



V. Model Recommendation

Between the three models tested in this analysis, I recommend the fourth-degree polynomial model. The trend line is much closer to a one-to-one ratio than the simple linear regression and marginally closer to one-to-one than the second-degree polynomial.

In all sets of predictions, there is a spike in data around 35 months of tenure. This is due to the underlying data, where a majority of customers have a tenure of around 35 months. This can be seen in the graph below.



VI. Summary

The main objective of this project was to develop a predictive model for churn, measured by how many months the customer was with the credit card company. The data set from Goyal (2020) included credit card customer tenure, gender, income and many other features. I cleaned the data using binomial, ordinal and one-hot-encoding transformations.

I analyzed the predicted values using a simple linear regression model and two different polynomial regression models. Though I planned to test a regularization model, the first polynomial model returned promising results and I decided a higher-order polynomial could return the best predictions for this data set.

VII. Next Steps

The improvement of the polynomial regression models with higher order should be further explored. Adding a cross-validation method, instead of the simple test-train split, can help identify the optimal degree for churn prediction. I will continue this effort and analysis with the media market measurement problem required for my job.