Analysis of Bike Buyers

Rita Philavanh
April 2017

Executive Summary

Analysis was performed on customer data collected by a bike retail company, Adventure Works Cycles (AWC). The dataset consists of 18361 total records, with each record containing details about a specific customer's demographics and their bike purchasing history. However, 6 of these records were found to be duplicated entries and the dataset was subsequently reduced to 18355 unique customer records.

To understand more about the patterns within the dataset, summary statistics and visualization was implemented on the cleansed data. This initial exploration identified potential relationships between customer demographic features and bike sales. Afterwards, the data was further analyzed by machine learning models. Classification models were used to classify the likelihood of a customer purchasing a bike (**BikeBuyer**), while regression models were used to predict a customer's average monthly spending with the company (**AvgMonthSpend**). These models were trained and tested to optimize their predictive accuracy and decrease their root mean squared error.

The results from the analysis led to discovering the following key findings:

- **Age** & **Gender**: Males between 30 to 50 years old have a distinctively higher Average Monthly Spend. The maximum **AvgMonthSpend** for the other age & gender demographics doesn't approach above 60. However, Males 30-50 show consistent patterns of **AvgMonthSpend** above 60.
- **Occupation/Education/Yearly Income**: There is a visible correlation between **Occupation**, **Education** and **YearlyIncome**. An order can be assumed with **Occupation** and **Education** that correlates with an increased **YearlyIncome**. It is also noted that as **YearlyIncome** increases, so too does the median **AvgMonthSpend**.
- **NumberChildrenAtHome**: Non bike buyers are significantly more likely to have no children at home. However, some bike buyers also have no children at home.
- **NumberCarsOwned**: Customer's median **AvgMonthSpend** values increased with the number of cars owned above 1. Their minimum value spent also noticeably increased as the number of cars owned increased.

Based on these results, it would be recommended to invest in marketing that targets males' ages between 30 and 50. While there were slight relationships observed in the other features, this demographic feature is most pronounced.

Initial Data Exploration

Before calculating summary statistics, data cleaning and feature engineering was performed in Python 3.6.0 to ensure all data is in their proper format for statistical analysis. This included removing duplicate customer records, handling problematic null values if they exist (they didn't), and converting textual categorical features into either numeric category labels (e.g., Education) or independent boolean features if there is no inherent order to the categories (e.g., Gender, MaritalStatus). New features were also derived as detailed in the table below:

Original Field	New Field	Purpose
BirthDate	Age	Easier to handle numeric values rather than dates
NumberChildrenAtHome	HasChildrenAtHome	Ability to investigate effect of having children overall
NumberCarsOwned	HasCars	Ability to investigate effect of owning more than 1 car

Individual Feature Statistics

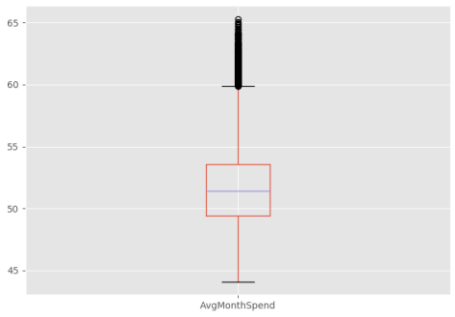
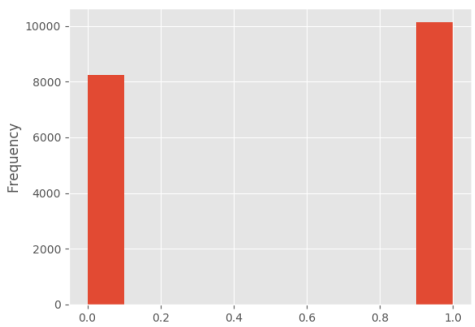
Summary statistics for minimum, maximum, mean, median, standard deviation, and percentiles were calculated for numeric features in order to get a feel for their distributions:

	#CarsOwned	#ChildrenAtHome	TotalChildren	YearlyIncome	Age
mean	1.27039	0.338218	0.850449	72758.95004	34.58
std	0.913887	0.569001	0.927363	30687.66436	11.26
min	0	0	0	25435	16
25%	1	0	0	53312.5	26
50%	1	0	0	61851	33
75%	2	1	2	87412	42
max	5	3	3	139115	86

For the boolean features, a grouped count of bike buyers was computed to get a feel for the bike buyer demographics. The statistics show home owners and married customers were more likely to be bike buyers at AWC than their counterparts:

		Number of BikeBuyers
HomeOwnerFlag	**0**	2924
	1	7203
Gender	**F**	4474
	M	5653
MaritalStatus	**M**	6348
	S	3779

AvgMonthSpend and **BikeBuyers** are labels we are interested in predicting, so we also investigate them through summary statistics and visualization:

Summary statistics of AvgMonthSpend	Box Plot of AvgMonthSpend	Histogram of Bike Buyers
count 18355 mean 51.767207 std 3.438024 min 44.100000 25% 49.410000 50% 51.420000 75% 53.600000 max 65.290000		

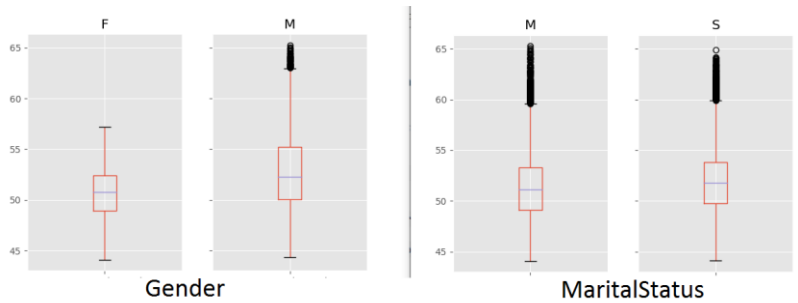
The box plot for **AvgMonthSpend** shows a great number of outlier points above 60. We will examine the demographic features further to find that certain age and gender explains this pattern.

Relationship between Age & Gender with AvgMonthSpend

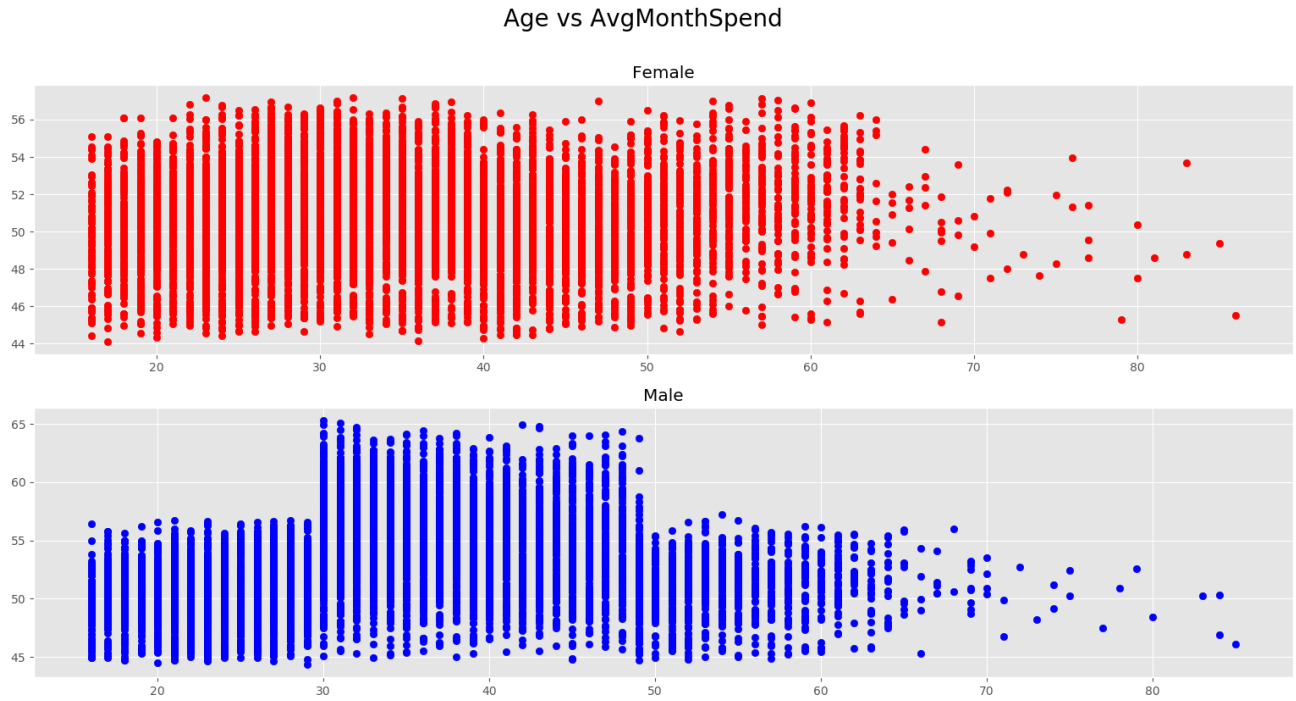
Calculating the summary statistics of the **AvgMonthSpend** for the boolean features, we discover Males have a higher upper **AvgMonthSpend**:

	HomeOwnerFlag		Gender		MaritalStatus	
	0	**1**	**F**	**M**	**M**	**S**
count	7148	11207	9070	9285	9945	8410
mean	50.51553	52.56555	50.68648	52.82291	51.54489	52.0301
std	2.937057	3.495884	2.480824	3.885862	3.522825	3.315992
min	44.1	44.48	44.1	44.33	44.1	44.16
25%	48.53	50.29	48.91	50.08	49.08	49.79
50%	50.26	52.09	50.76	52.28	51.15	51.78
75%	52.3	54.47	52.39	55.24	53.29	53.83
max	63.94	65.29	57.19	65.29	65.29	64.91

This pattern can also be clearly seen in the box plot for **Gender** (and as comparison, the pattern did not exist for **MaritalStatus)**:



After investigating other numerical features in the data, it is found to be a distinct pattern for males aged 30-50 having a higher **AvgMonthSpend**:



So we can conclude that **Gender** and **Age** appear to be key features for **AvgMonthSpend**.

Relationship between Occupation, Education, YearlyIncome with AvgMonthSpend

Next, other categorical features were examined, such as **Occupation** and **Education**.

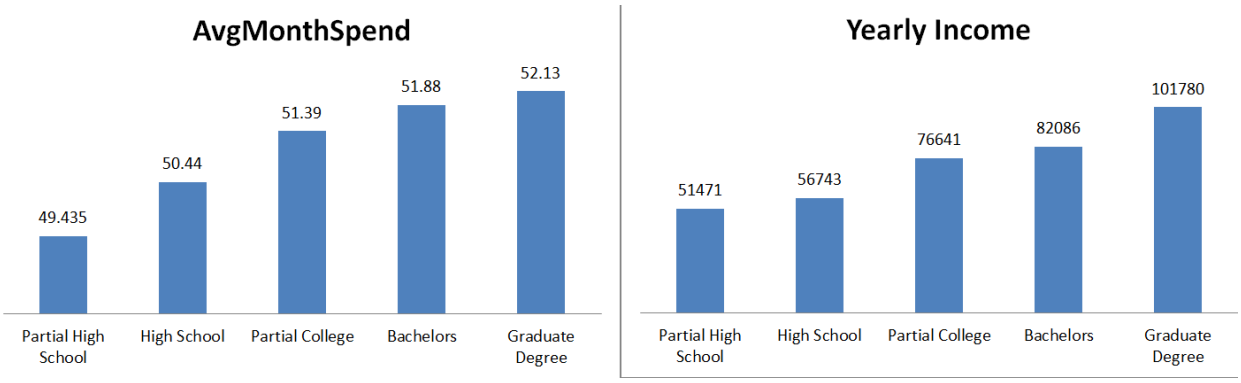
Bike Buyers

	1	0		1	0
Manual	886	2489	**Partial High School**	471	1103
Skilled Manual	3386	2672	**High School**	1708	1566
Clerical	2849	1612	**Partial College**	3164	2157
Management	1913	945	**Bachelors**	2838	2185
Professional	1093	510	**Graduate Degree**	1946	1217

No patterns appear to exist with **BikeBuyer**, however, a correlation appears to exists with the ordered **Education** categories, the **YearlyIncome** and median **AvgMonthSpend**:

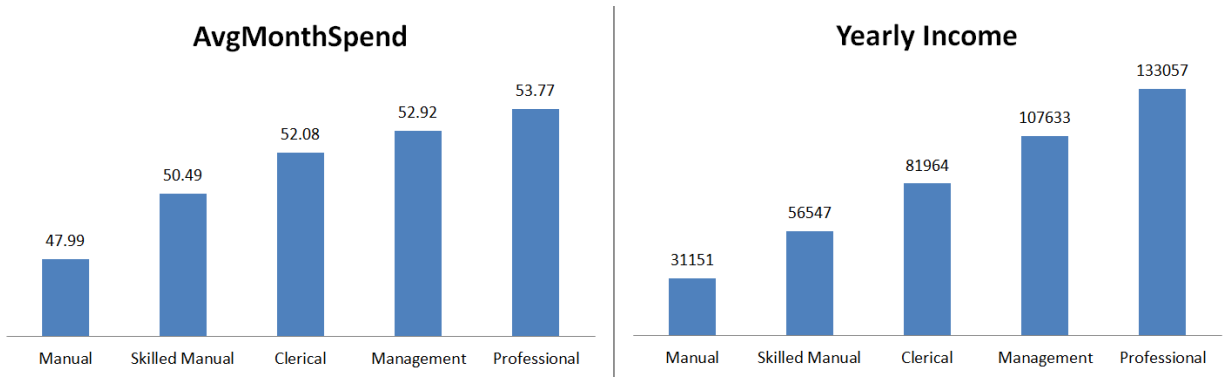
Education	Median Yearly Income	Median AvgMonthSpend
Partial High School	51471	49.435
High School	56743	50.44
Partial College	76641	51.39
Bachelors	82086	51.88
Graduate Degree	101780	52.13

This correlation becomes more apparent when plotted:



This same correlation exists for an ordered **Occupation** category:

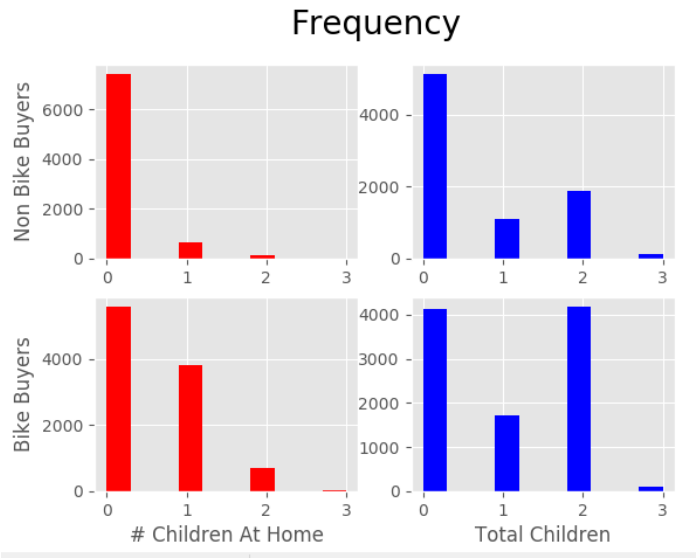
Occupation	Median Yearly Income	Median AvgMonthSpend
Manual	31151	47.99
Skilled Manual	56547	50.49
Clerical	81964	52.08
Management	107633	52.92
Professional	133057	53.77



Based on this, we can assume a categorical ordering for **Education** and **Occupation** that correlates with **YearlyIncome,** and as a result, with **AvgMonthSpend**.

Relationship between NumberChildrenAtHome with BikeBuyer

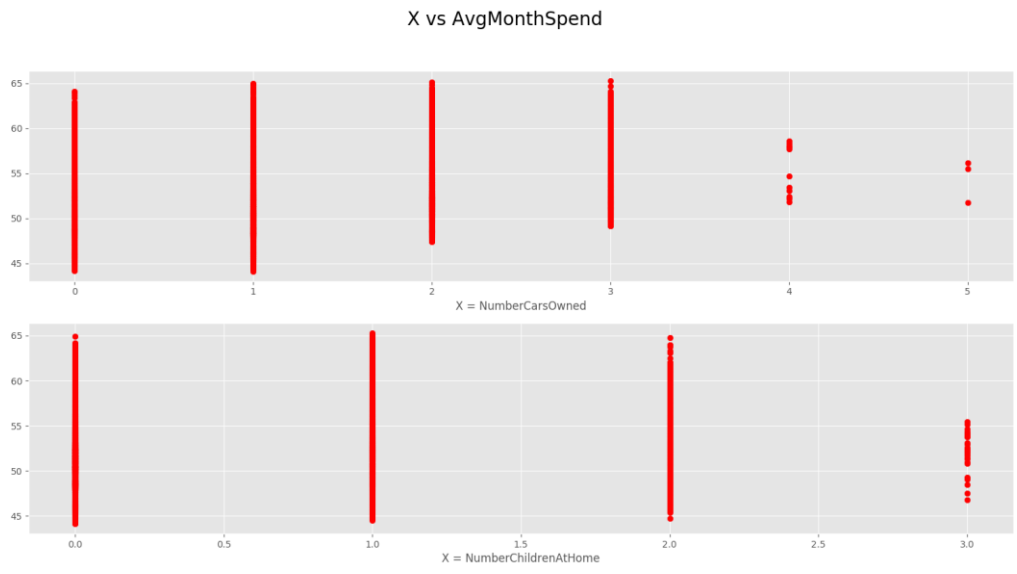
Plotting a histogram of bike buyers grouped by number of children at home (in red below), it was found that non bike buyers are more likely to have no children at home. The reverse wasn't true since many bike buyers have no children. But if they were to have children at home, they are more likely to be bike buyers. This trend wasn't as pronounced in the histograms for Total Children (in blue). In this latter case, there were several customers who had children but were non bike buyers.



From this, we can conclude that if a customer has children *at home*, they are slightly more likely to be bike buyers. However, just because a customer has no children doesn't mean they are non bike buyers.

Relationship of NumberCarsOwned with AvgMonthSpend

Taking a look at another feature, **NumberCarsOwned**, we see an increase in **AvgMonthSpend** with the number of cars owned above 1. The minimum spend is also noticeably increasing as seen in this plot (and as comparison, this trend was not observed in **NumberChildrenAtHome**):



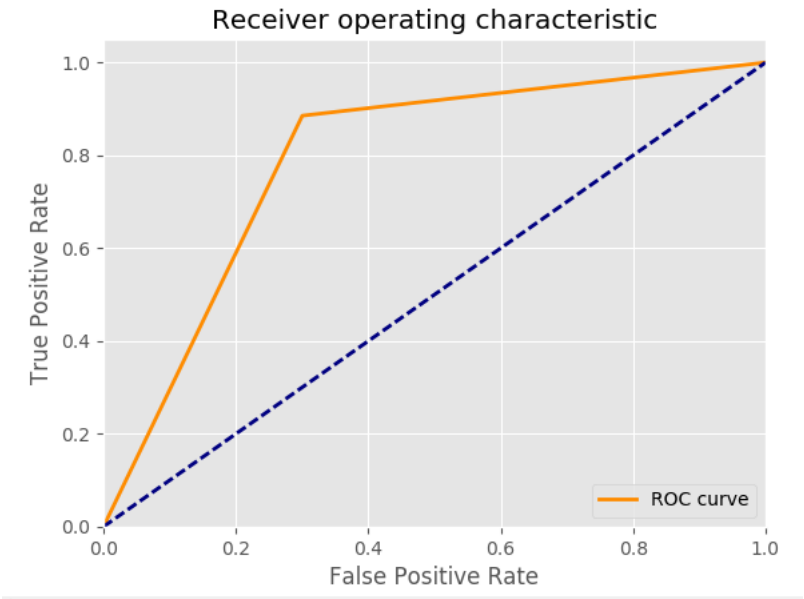
After examining the key demographic features and deriving new features, we are ready to begin predictive analysis on the dataset. All non categorical text features (e.g. Name, Address) will be dropped from the feature list before trainig the models.

Classification of Customers as being BikeBuyers

A Two Class Decision Tree Classifier algorithm with max depth of 6 was trained with 70% of the data. Testing the model with the remaining 30% of the data led to the following results:

- True Positives: 1754
- True Negatives: 2656
- False Positives: 753
- False Negatives: 344

A Received Operator Characteristic (ROC) curve for the model is plotted below. This plot shows the performance of a model at various classification thresholds of True Positive vs False Positive rates:

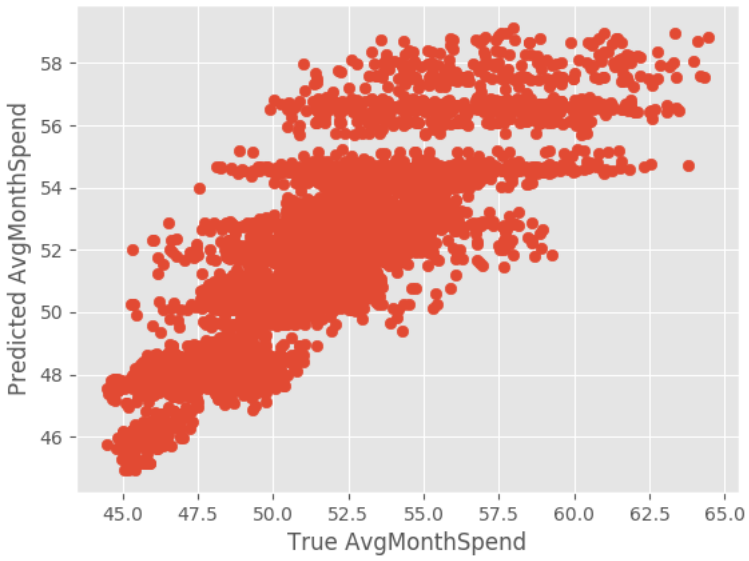


Since the diagonal line is the expected results of a random guess, the orange line for the model's performance indicates better classification than random guess. This means that the model has some predictive capabilities. Calculating the exact metrics of the model's performance:

- Accuracy: 80.08%
- Precision: 77.91%
- Recall: 88.53%
- F1 Score: 82.88%

Regression of AvgMonthSpend

After creating a classification model to predict bike buying probability, a regression model was created to predict **AvgMonthSpend** of a customer. A Gradient Boosting Regressor was trained with 70% of the data then tested with the remaining 30% of the data. The results are depicted in a scatter plot below:



This plot shows a linear relationship between predicted and actual values in the test dataset. The Root Mean Square Error (RMSE) for the test results is 1.96.

Conclusion

This analysis has shown that the probability of a customer being bike buyers with AWC or not can be predicted from the customer's demographic features with good accuracy. Home ownership, marital status and number of children at home were features that can help classify a customer as being bike buyers or not. Further, the **AvgMonthSpend** value could also be estimated with low error. Age, gender and Yearly income are some of the features identified as having a significant influence on a customer's **AvgMonthSpend**.