Adventure Works Customer Prediction Analysis

Team Zebra

Presentation Outline

1. Data Cleaning
2. Data Visualization
3. Predictive Models
4. Conclusions and Action Items

Data Cleaning

- Visual sanity check
- Remove duplicate values and check for errors in the data
- Discard useless features (Zip Code, Title, Name, etc)
- Check for outliers
- Compute any necessary data such as age
- Join the datasets

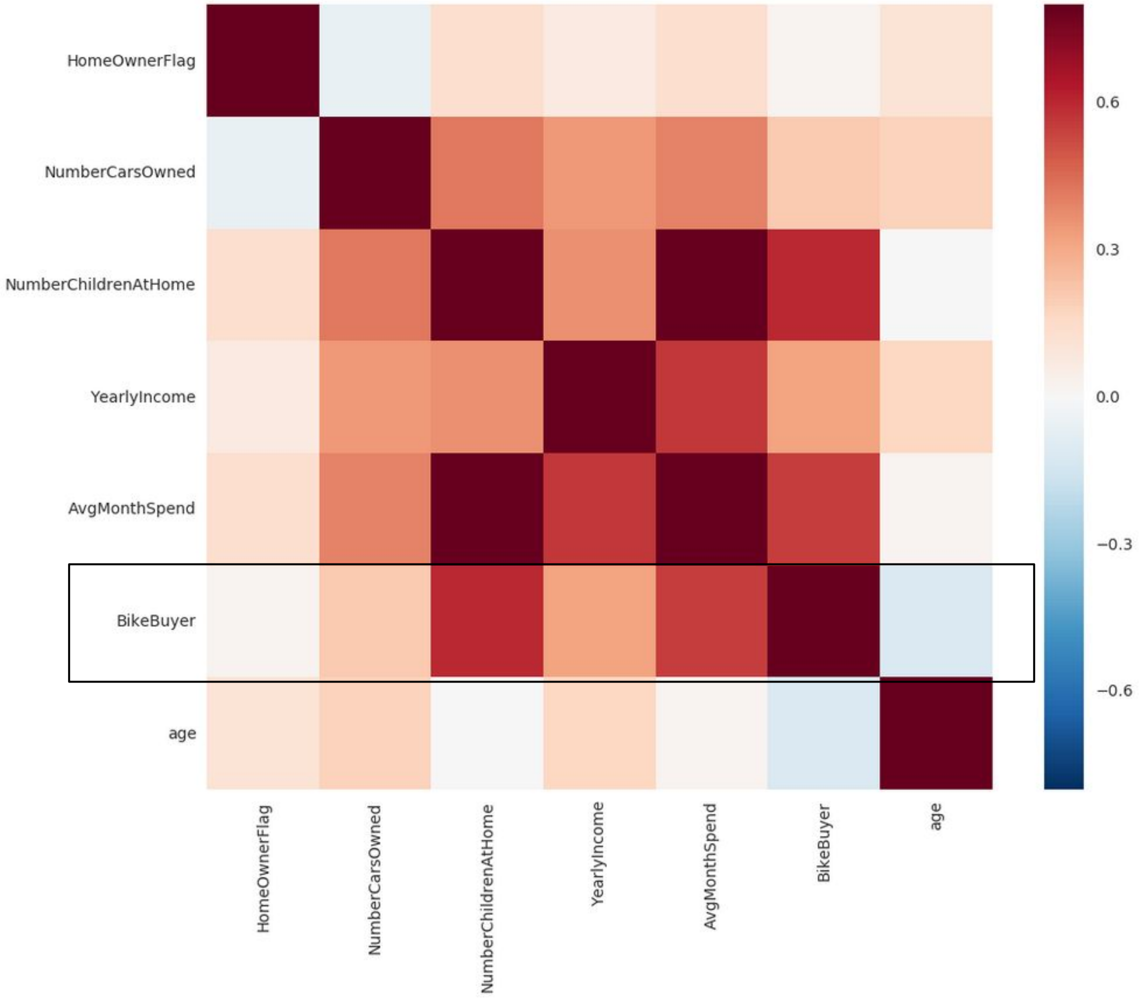
Correlations

Bike Buyers highly correlated with:

- Average monthly spending
- Number of children at home

Average Monthly Spending highly correlated with:

- Number of children at home
- Yearly Income
- Average Monthly Spending



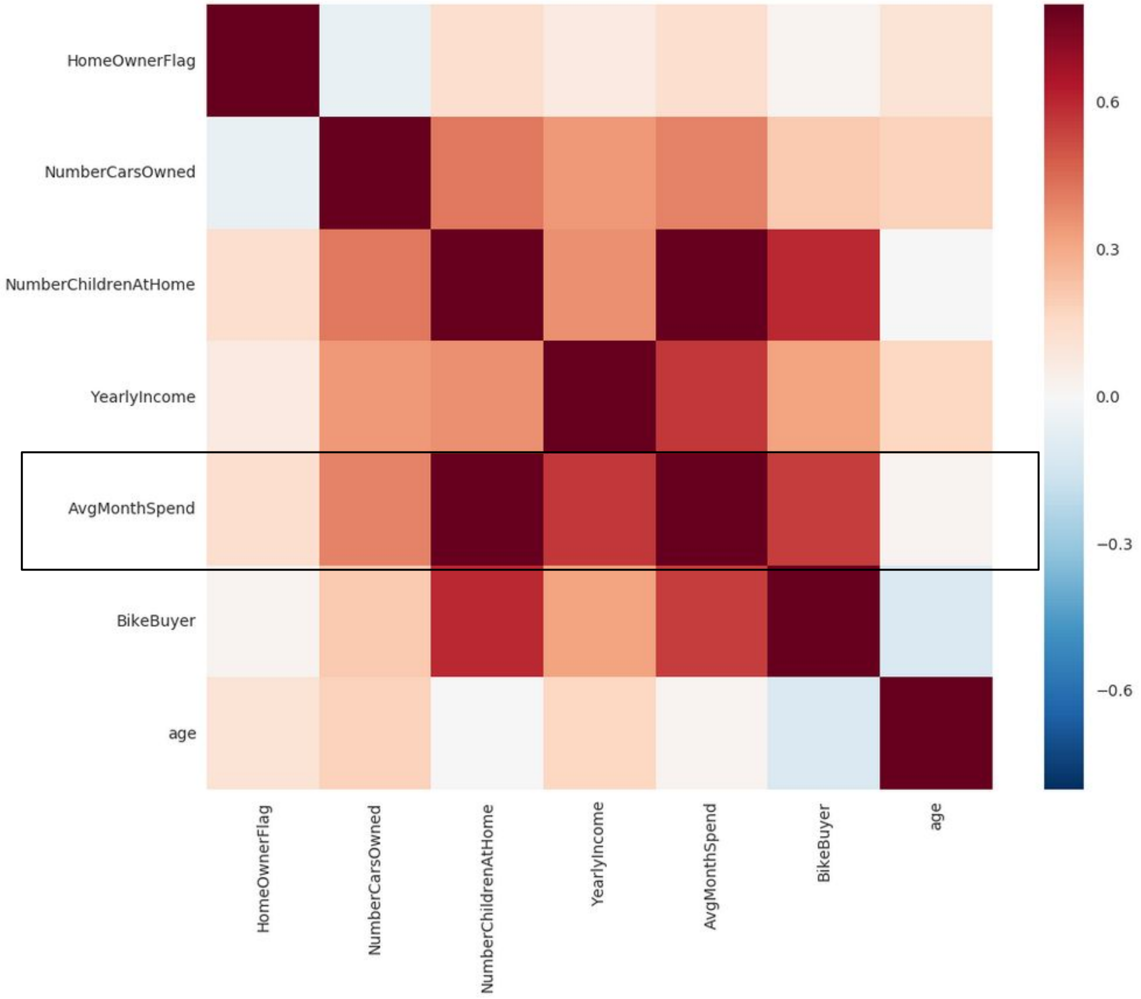
Correlations

Bike Buyers highly correlated with:

- Average monthly spending
- Number of children at home

Average Monthly Spending highly correlated with:

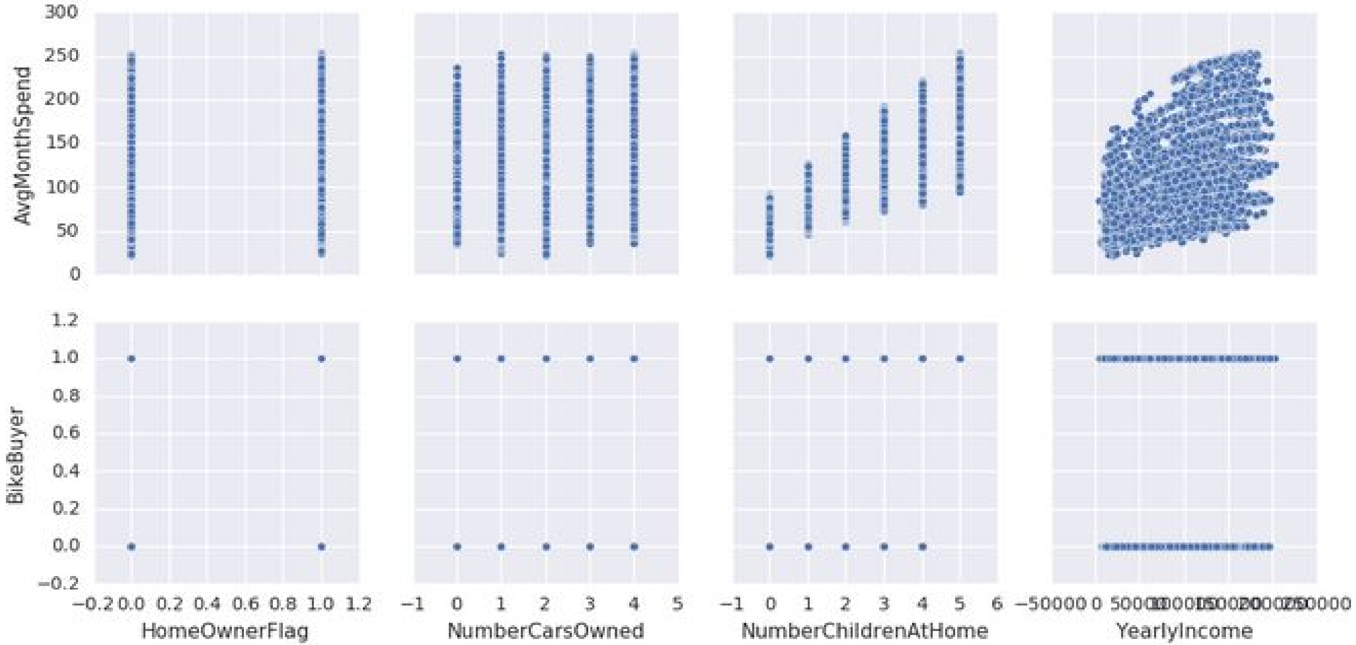
- Number of children at home
- Yearly Income
- Average Monthly Spending



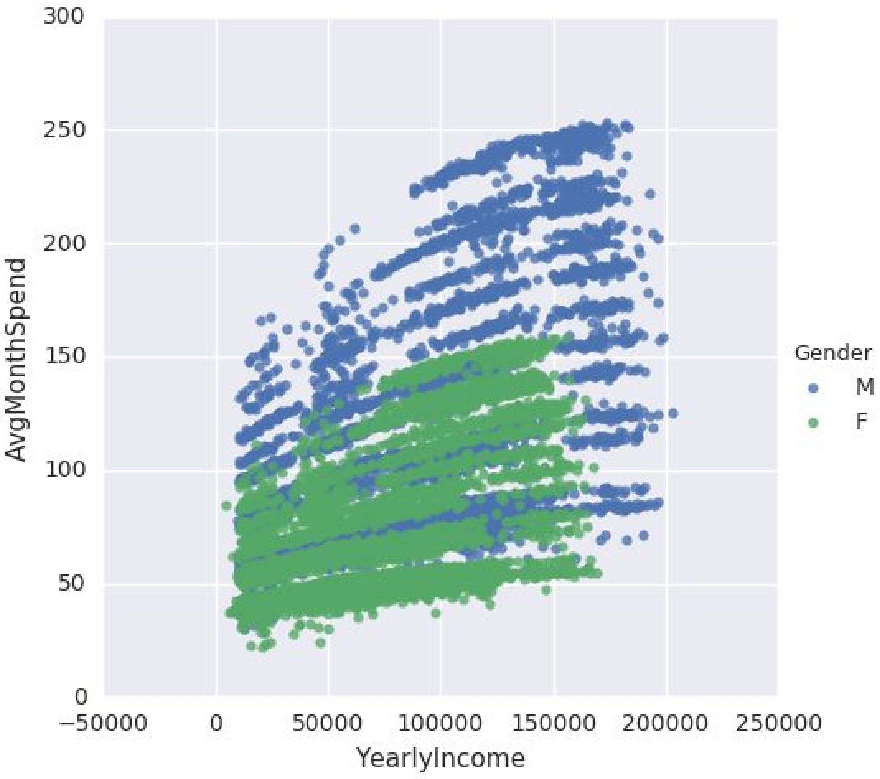
Data Exploration

Correlations between numerical outcomes.

Data Exploration

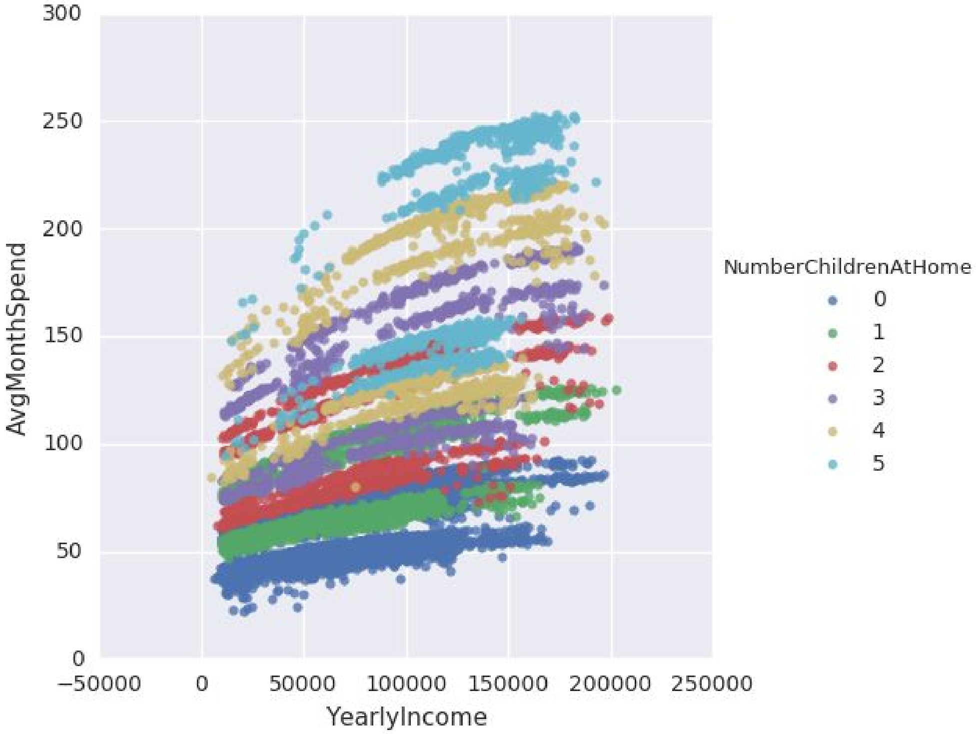


Gender

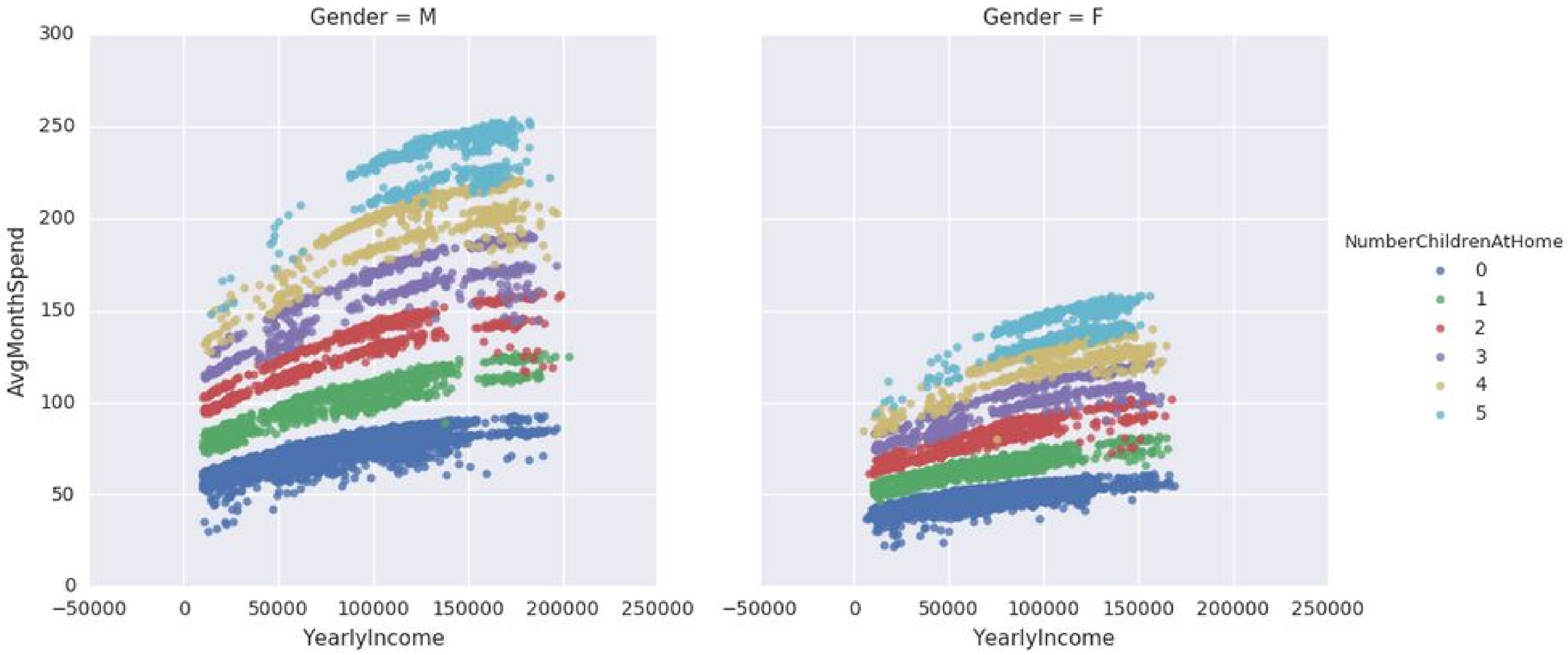


300
250
200
150
100
50
0
AvgMonthSpend
−50000
0
50000
100000
150000
200000
250000
YearlyIncome
Gender
M
F

Number of Children



Customer Demographics



Create An Indicator Variable With R-Script

Variable = Income * isMale/Female * isChildren at Home?

```r
# Generate 12 Income Labels For Specific Groups
m_zero_inc <- ifelse(dataset1$Gender == 'M' & dataset1$NumberChildrenAtHome == '0',dataset1$YearlyIncome,0)
m_one_inc <- ifelse(dataset1$Gender == 'M' & dataset1$NumberChildrenAtHome == '1',dataset1$YearlyIncome,0)
m_two_inc <- ifelse(dataset1$Gender == 'M' & dataset1$NumberChildrenAtHome == '2',dataset1$YearlyIncome,0)
m_three_inc <- ifelse(dataset1$Gender == 'M' & dataset1$NumberChildrenAtHome == '3',dataset1$YearlyIncome,0)
m_four_inc <- ifelse(dataset1$Gender == 'M' & dataset1$NumberChildrenAtHome == '4',dataset1$YearlyIncome,0)
m_five_inc <- ifelse(dataset1$Gender == 'M' & dataset1$NumberChildrenAtHome == '5',dataset1$YearlyIncome,0)

f_zero_inc <- ifelse(dataset1$Gender == 'F' & dataset1$NumberChildrenAtHome == '0',dataset1$YearlyIncome,0)
f_one_inc <- ifelse(dataset1$Gender == 'F' & dataset1$NumberChildrenAtHome == '1',dataset1$YearlyIncome,0)
f_two_inc <- ifelse(dataset1$Gender == 'F' & dataset1$NumberChildrenAtHome == '2',dataset1$YearlyIncome,0)
f_three_inc <- ifelse(dataset1$Gender == 'F' & dataset1$NumberChildrenAtHome == '3',dataset1$YearlyIncome,0)
f_four_inc <- ifelse(dataset1$Gender == 'F' & dataset1$NumberChildrenAtHome == '4',dataset1$YearlyIncome,0)
f_five_inc <- ifelse(dataset1$Gender == 'F' & dataset1$NumberChildrenAtHome == '5',dataset1$YearlyIncome,0)
```

*Allows Different Slopes for Groups In Previous Slide

Groups of Predictors

3 Key Predictive Features

- Gender, Number of Children, Income- Interaction Term (*From Previous Slide)

9 Less Predictive Features

- Marital Status, Number of Cars, City, State/Province, Country/Region, Education, Occupation, HomeOwner, Age

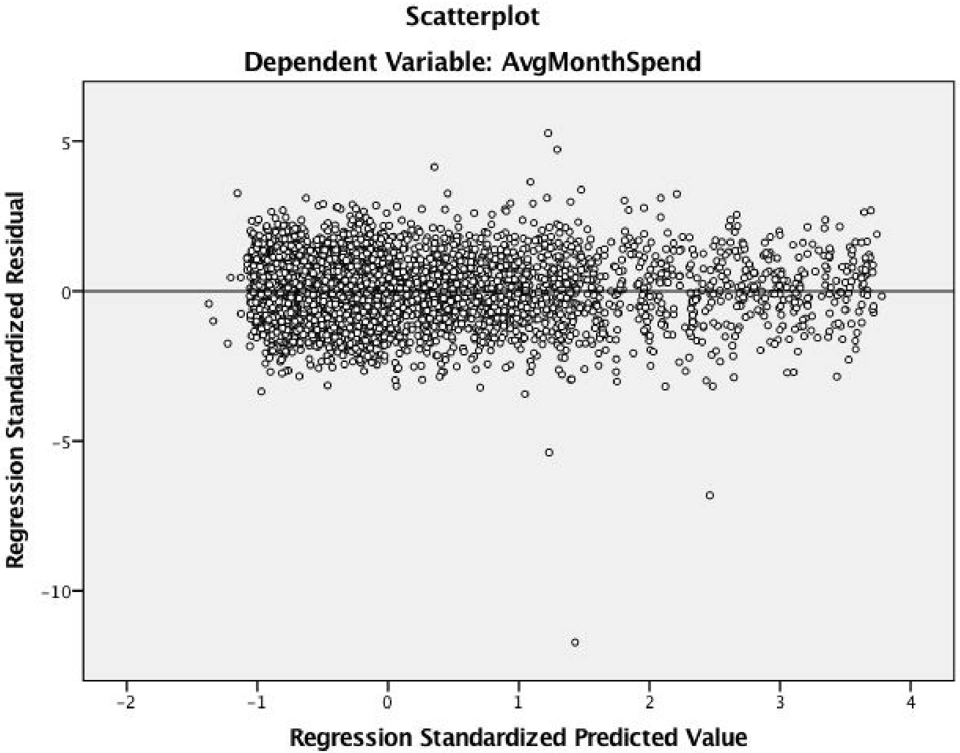
Importance of Key Features in Model

Nested Model Without Key Features:

Full Model Including Key Features:

Metrics (Nested Model)		Metrics (Full Model)	
Mean Absolute Error	19.745669	Mean Absolute Error	0.939415
Root Mean Squared Error	25.897131	Root Mean Squared Error	1.22055
Relative Absolute Error	0.60587	Relative Absolute Error	0.028825
Relative Squared Error	0.33938	Relative Squared Error	0.000754
Coefficient of Determination	0.66062	Coefficient of Determination	0.999246

Residual Plot From Prediction Algorithm: Only Key Features Used



Classification of Bike Buyers

Nested Model Without Key Features

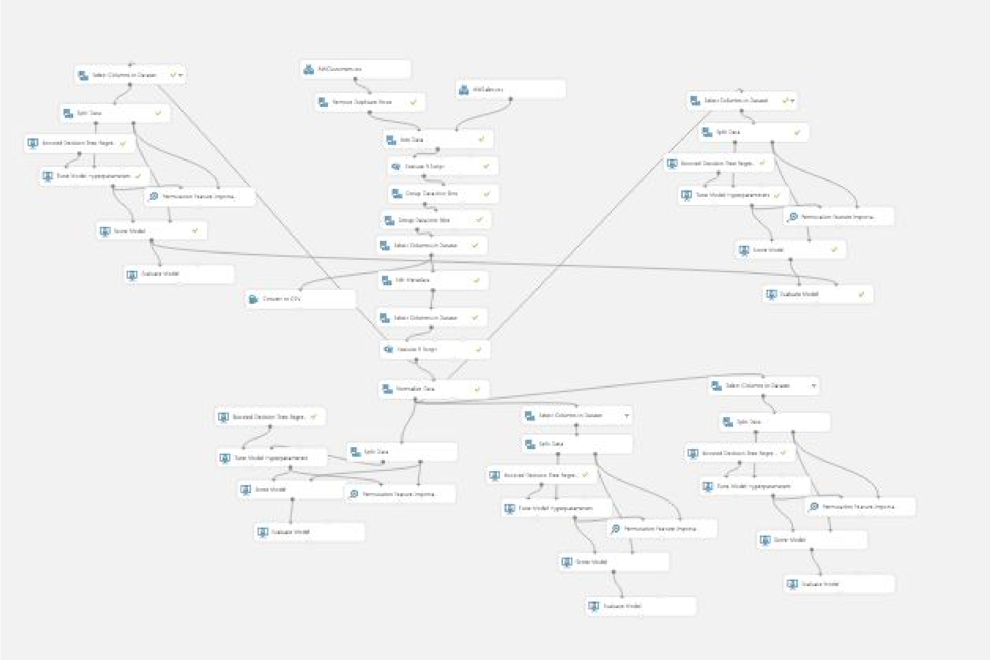
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1584	657	0.779	0.739	0.5	0.850
False Positive	**True Negative**	**Recall**	**F1 Score**		
560	2714	0.707	0.722		

Full Model With Key Features

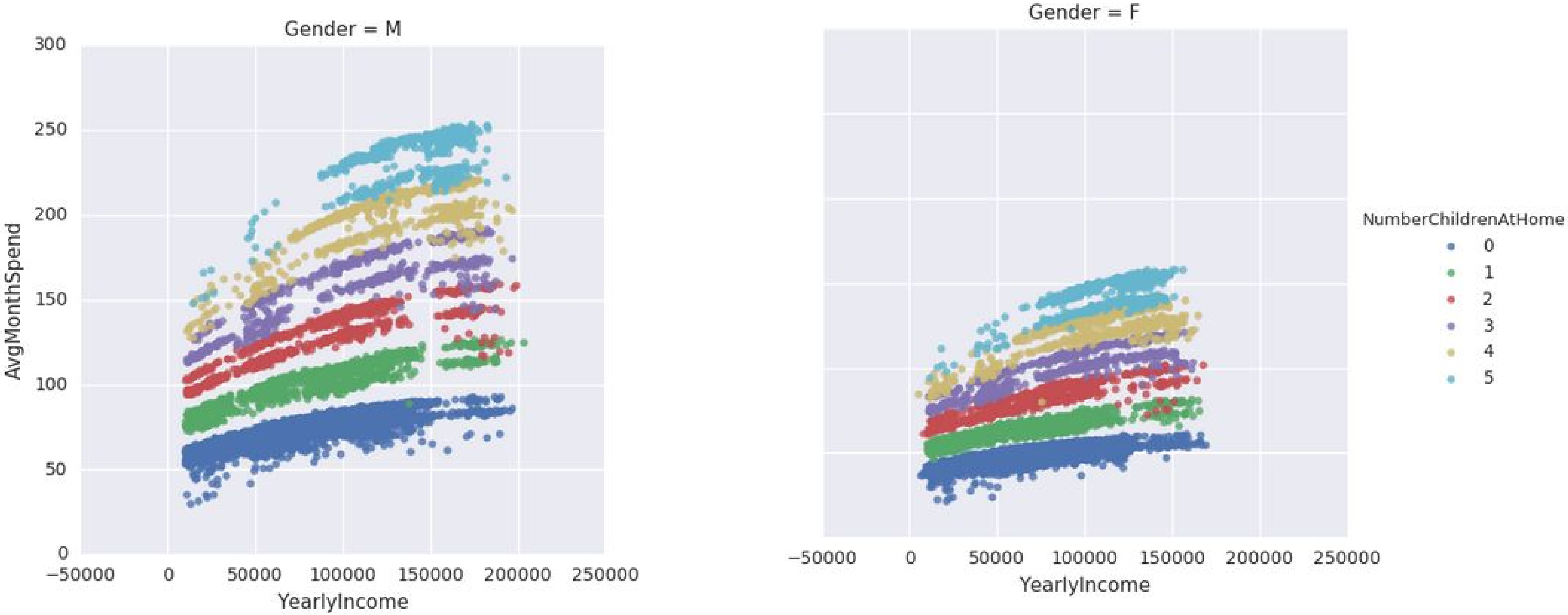
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1670	571	0.805	0.768	0.5	0.885
False Positive	**True Negative**	**Recall**	**F1 Score**		
505	2769	0.745	0.756		

Algorithm Selection Process

Compare Models In Parallel On Same Features



Conclusions



- Males who have higher income will spend more
- Within the male group, spending increases with more children
- We see the same trends with females, although females spend less overall
- Bottom line - we want to target males with higher income and more children. However, there is a sweet spot where we should switch and target females