ANALYTIXLABS

Sample ML Project Outputs

Date: 11th March 2019

Agenda

- ✓ **Business Problem & Goals**
- ✓ **Project Approach**
- ✓ **Solution Approach**
- ✓ **Exploratory Data Analysis**
- ✓ **Data Assumptions**
- ✓ **Data Understanding & Data Audit**
- ✓ **Univariate Analysis**
- ✓ **Bivariate Analysis**
- ✓ **Data Preparation & Feature Reduction**
- ✓ **Model Building & Fine Tuning**
- ✓ **Model Validation Outputs**
- ✓ **Recommendations**

ANALYTIXLABS

Business problem

- **Business Context:** One of the global banks would like to understand what factors driving credit card spend are. The bank want use these insights to calculate credit limit.

- **Business Goals:** Survey data is available for 5000 customers
  - Predict Total Spend for customers (total of primary and secondary card spends)
  - Establish credit limits based on the key drivers
  - Given the factors, predict credit limit for the new applicants

- **Approach:**
  - Carry out regression to determine total spend
  - Establish relationship to determine credit limit.

Project Approach

Objective
- **Establish factors driving credit card spend**
- **Predict Credit card spent (primary +secondary cards)**
- **Determine credit limits**

W1 Business Problem & DATA Audit

- Import Data
- Data Profiling – Types, distribution, classification (Ordinal/Nominal/Continuous)
- Data statistics
  - Min, max, unique, missing values, outliers etc
  - IQRs, Outliers
  - Skewness
- Correlations
  - Categorical vs continuous
  - Continuous vs continuous
- Transformation/Standardization of features
- F-Regression/VIF

Outputs
Approach & summary
Profiling
Data statistics

W2 First Model

- Feature Engineering
  - Missing value treatments
  - Outlier treatment
  - Feature transformations – alternate approaches
  - Dimensionality reduction - PCA
- Model building
  - Train/test splits
  - Select model
  - Model fit for train & test
  - Metrics for train & test
- Business outputs
  - Excel workbook with model simulation

Outputs
Model parameters
Metrics
Output for base model

W3 Model Fine-tuning

- Try different models
- Revisit Feature engineering
- Hyper tuning
- Metrics comparison

Outputs
Metrics comparison between models
Finalize model

W4 Final Outputs

- Create Outputs for submission
- Presentation
  - Process
  - Observations
  - Comparisons
  - Conclusions
- Artefacts
  - Detailed Code
  - Detailed outputs
  - Interim outputs
- Model simulation workbook

Outputs
Final Presentation
Detailed Excel outputs
Detailed codes

Solution Approach

Data Preparation

- Import Data
- EDA - Statistics
- Preprocess: Creation of columns, Outlier treatment, Missing values
- EDA: Calculate feature statistics, distribution relationships such as R2, Skew,

Data processing

- PCA/VIF/Factor analysis
- MODEL SELECTION: Try different models – Linear Regression, Random Forest, Gradient Boost, XG Boost. Also combination of the above
- Hyper parameter tuning. Create model objects (like pickles). To be used by the Business Output function

Business Objectives

Credit Card Spend

- XGBoost – Best estimator (after multiple iterations) for predicting total spend
- Credit limit based on the most influential drivers – Income, Debt ratio etc

Exploratory Data Analysis

Data Related Assumptions

- Target variable:- Create a total spent column by adding card spend on Primary and Secondary card.
- There are variables ("carown", "cartype", "carcatvalue" , "carbought") which have -1 as the value. This will be treated as missing value and imputed accordingly
- Missing's less than 20% will be imputed with mode and constant 0(zero)
- As the data contains categorical conversion of some of the numerical variables, we will consider the categorical versions of the same for exploratory analysis.

ANALYTIXLABS

Data Understanding

- **Input Data:**
  Total : 5000+ customers.
  Total Variables = 130+
  84 variables are categorical and 36 are continuous variables

  ('custid', 'region', 'townsize', 'gender', 'age', 'agecat', 'birthmonth', 'ed', 'edcat', 'jobcat', 'union', 'employ', 'empcat', 'retire', 'income', 'lninc', 'inccat', 'debtinc', 'creddebt', 'lncreddebt', 'othdebt', 'lnothdebt', 'default', 'jobsat', 'marital', 'spoused', 'spousedcat', 'reside', 'pets', 'pets_cats', 'pets_dogs', 'pets_birds', 'pets_reptiles', 'pets_small', 'pets_saltfish', 'pets_freshfish', 'homeown', 'hometype', 'address', 'addresscat', 'cars', 'carown', 'cartype', 'carvalue', 'carcatvalue', 'carbought', 'carbuy', 'commute', 'commutecat', 'commutetime', 'commutecar', 'commutemotorcycle', 'commutecarpool', 'commutebus', 'commuterail', 'commutepublic', 'commutebike', 'commutewalk', 'commutenonmotor', 'telecommute', 'reason', 'polview', 'polparty', 'polcontrib', 'vote', 'card', 'cardtype', 'cardbenefit', 'cardfee', 'cardtenure', 'cardtenurecat', 'card2', 'card2type', 'card2benefit', 'card2fee', 'card2tenure', 'card2tenurecat', 'cardspent', 'card2spent', 'active', 'bfast', 'tenure', 'churn', 'longmon', 'lnlongmon', 'longten', 'lnlongten', 'tollfree', 'tollmon', 'lntollmon', 'tollten', 'lntollten', 'equip', 'equipmon', 'lnequipmon', 'equipten', 'lnequipten', 'callcard', 'cardmon', 'lncardmon', 'cardten', 'lncardten', 'wireless', 'wiremon', 'lnwiremon', 'wireten', 'lnwireten', 'multline', 'voice', 'pager', 'internet', 'callid', 'callwait', 'forward', 'confer', 'ebill', 'owntv', 'hourstv', 'ownvcr', 'owndvd', 'owncd', 'ownpda', 'ownpc', 'ownipod', 'owngame', 'ownfax', 'news', 'response_01', 'response_02', 'response_03')

  Missing info columns:
  - >20% (lntollmon, lntollten, lnequipmon, lnequipten, lncardmon, lncardten, lnwiremon, lnwireten)
  - <20% (townsize, lncreddebt, lnothdebt, commutetime, longten, lnlongten, cardten)

ANALYTIXLABS

Exploratory Data Analysis

User created

Snapshot of some important features

Feature	N	NMISS	Unique	SUM	MEAN	MEDIAN	STD	VAR	skew	skewtest_	skewtest_	IQRLowOu	IQRHighO	corelation	pvalcards	corelation	pvalcard2	corelation	pvalTotals	TotalCorrelation
TotalSpend	5000	0	4886	2490393	498.0786	414.25	351.5293	123572.8	2.707547	45.22287	0	0	264	0.939815	0	0.829933	0	1	0	2.769747975
cardspent	5000	0	4760	1686013	337.2025	276.36	245.1451	60096.1	3.051237	47.94182	0	0	256	1	0	0.597256	0	0.939815	0	2.537071101
card2spent	5000	0	4477	804380.6	160.8761	125.34	146.2928	21401.57	2.801163	45.99372	0	0	248	0.597256	0	1	0	0.829933	0	2.427189397
SpendRatio	5000	0	4994	66115.23	13.22305	10.5066	10.60701	112.5087	2.426434	42.75254	0	0	218	0.369521	1.40E-161	0.330444	1.10E-127	0.393581	6.20E-185	1.093545847
income_cat	5000	0	12	20901	4.1802	4	1.644223	2.703469	1.46666	31.85411	1.15E-222	0	153	0.358217	2.82E-151	0.318061	6.29E-118	0.380098	1.32E-171	1.056375652
income	5000	0	266	273798	54.7596	38	55.37751	3066.669	5.179216	60.19145	0	0	346	0.355016	1.96E-148	0.318498	2.89E-118	0.3772	8.06E-169	1.050714307
carvalue	5000	0	801	116162.9	23.23258	17	21.23164	450.7824	1.474026	31.95767	4.23E-224	0	334	0.302148	4.70E-106	0.266705	3.49E-82	0.319367	6.17E-119	0.888220334
carvalue_cat	5000	0	5	12590	2.518	2	1.335677	1.784033	0.506401	13.83406	1.59E-43	0	0	0.310966	1.52E-112	0.254089	1.64E-74	0.320673	6.00E-120	0.885728892
CardSpentPercentage	5000	0	4816	344965.7	68.99313	69.60307	15.28502	233.6317	-0.30185	-8.53762	1.37E-17	33	0	0.143112	2.72E-24	-0.56705	0	-0.13767	1.38E-22	0.847837208
othdebt	5000	0	4973	18272.3	3.65446	2.09854	5.395172	29.10788	7.589924	69.1551	0	0	390	0.260531	2.24E-78	0.227568	9.59E-60	0.272972	3.75E-86	0.761070977
othdebt_cat	5000	0	7	8351	1.6702	1	1.031138	1.063245	1.545902	32.94781	4.55E-238	0	353	0.248667	2.41E-71	0.214769	2.93E-53	0.260061	4.32E-78	0.723496096
creddebt	5000	0	4950	9286.628	1.857326	0.926437	3.415732	11.66722	11.12004	78.15443	0	0	443	0.233253	9.29E-63	0.21089	2.25E-51	0.248547	2.82E-71	0.692690931
retire														-0.15816	2.28E-29	-0.14386	1.56E-24	-0.16933	1.77E-33	0.471347605
IncomeDebtRatio	5000	0	4998	70582.27	14.11645	11.33301	10.30952	106.2862	3.471181	50.90016	0	0	309	0.158518	1.70E-29	0.13604	4.37E-22	0.167677	7.48E-33	0.462234819

Feature	CV	MIN	MAX	P1	P5	P10	P25	P50	P75	P90	P95	P99	IQR
TotalSpend	0.705771	8.11	4881.05	58.1976	133.106	184.033	276.2825	414.25	615.5625	908.125	1145.147	1760.102	339.28
cardspent	0.726997	0	3926.41	40.8195	91.3045	122.534	183.3775	276.36	418.5375	610.062	782.3155	1215.807	235.16
card2spent	0.90935	0	2069.25	0	14.8195	28.639	66.9675	125.34	208.31	324.718	419.447	712.1298	141.3425
SpendRatio	0.802161	0.253939	127.2656	1.332091	2.654685	3.655484	5.975185	10.5066	17.19736	26.36023	32.55857	50.93873	11.22217
income_cat	0.393336	0	11	1	2	3	3	4	5	6	7	10	2
income	1.011284	9	1073	9	13	16	24	38	67	109.1	147	272.01	43
carvalue	0.913873	-1	99.6	-1	-1	2.49	9.2	17	31.1	52.91	72	92.001	21.9
carvalue_cat	0.530452	1	5	1	1	1	1	2	4	5	5	5	3
CardSpentPercentage	0.221544	0	100	31.78026	42.94276	49.38461	59.11253	69.60307	79.43506	88.17726	93.61525	100	20.32253
othdebt	1.476325	0	141.4592	0.114299	0.287692	0.457997	0.980301	2.09854	4.31478	8.062046	11.81598	24.06426	3.334479
othdebt_cat	0.617374	0	6	1	1	1	1	1	2	3	4	5	1
creddebt	1.839059	0	109.0726	0.03316	0.101088	0.175682	0.38552	0.926437	2.06382	4.29947	6.37301	14.28036	1.678301
retire													
IncomeDebtRatio	0.730319	2.52241	150	3.839107	5.095209	5.954887	7.805129	11.33301	16.81385	25.24347	32.87957	54.91886	9.008716

Univariate Analysis (X Variables – Numerical): Distributions

- Some of the variables are showing similar distribution to the Total Spent e.g. ed, debtinc, income, creddebt and othdebt
- There are a lot of n-modal distributions present e.g. address, age, cars, pets, reside, cardmon,cardten,equipmon,equipten, wiremon, wireten. For these, each mode will have to be addressed separately. Also check the contribution from each mode. One solution will be to convert them to categorical variable.
- There are a lot of zero values in some of the variables like cardmon, cardten, equipmon, equipten, othdebt, pets, tollmon, tollten, wiremon, wireten. It simply shows that these products or services are not used in last month by these customers. One solution will be to convert them to categorical variable.
- The information from pariplot of totalspent with other numerical variables shoes some linear relatioship with the variables income, creddebt and othdebt. Some other variables have linear relatioship but it is not that significant.

Data Audit Numerical graphs

Univariate Analysis (X Variables – Categorical): Distributions

- Some of the variables are evenly spread and do not clearly show distinction between the count distribution i.e they are evenly distributed e.g. region, gender, jobsat, marital, vote, cardtype, cardbenefit,bfast,callid, callwait, forward, confer, ownipod, owngame, news
- Some of the variables with binary categories are heavily skewed, e.g. union, retire, default, cardfee, owntv, ownvcr, owncd, owndvd, ownpda, ownfax, response_01, response_02, response_03
- Variables that have more than 2 categories also show skewed patterns, e.g. spousedcat, carown, commute, reason
- There are higher number of card users from townsize=1, But average overall spending from all the towns is same.
- Female cardholdes are slighty higher spenders than Male cardholders
- People having "Managerial and Professional" and "Sales and Office" are the top card holders. While people having jobs of "Managerial and Professional" and "Service" are highest spenders.
- Most of the cardholders are not retired and do not belong to any union.
- Married people hold less cards than unmarried people. But their average spending is similar.
- Average spending of people defaulting on bank loan is similar to ones that have not defaulted although the cardholders in the default category are very less
- The spending increases with the increase in job satisfaction level

Data Audit
riate and Bivariate

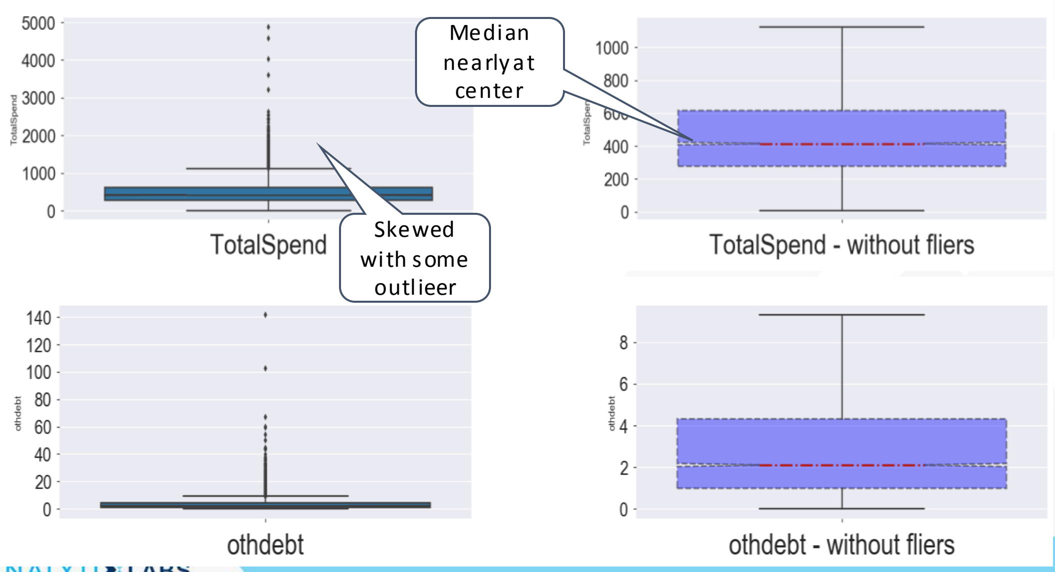
ANALYTIXLABS

Univariate Analysis (X Variables - Categorical: [Cont..]

- More cardholders own a house. But their average spending is similar to cardholders who rent a house.
- Single-Family cardholders are more likely to hold a card.
- Cars owners are higher number of cardholders compared to car leasing.
- There are variables ("carown", "cartype", "carcatvalue" , "carbought") with value as "-1" i.e. "N/A", for which we might have to perform missing value treatment.
- There is increase in spend from standard to Luxury car owners.
- Car commuting cardholders are the highest cardholders. But average spend among all type of commute is similar.
- Choosing a card does not depend significantly on any particular reason.
- Extremely conservative people are less spenders as well as they are less among all cardholders. All others have similar average spend.
- AMEX card users are highest average spenders but maximum cardholders are having VISA, MASTER and DISCOVER cards
- Lot of people prefer cards without fee
- Lot of cardholders do not have internet connection
- More cardholders prefer paper bill
- Overall spend of people who owns items like vcr, cd, dvd, tv is higher than those who do not own these items

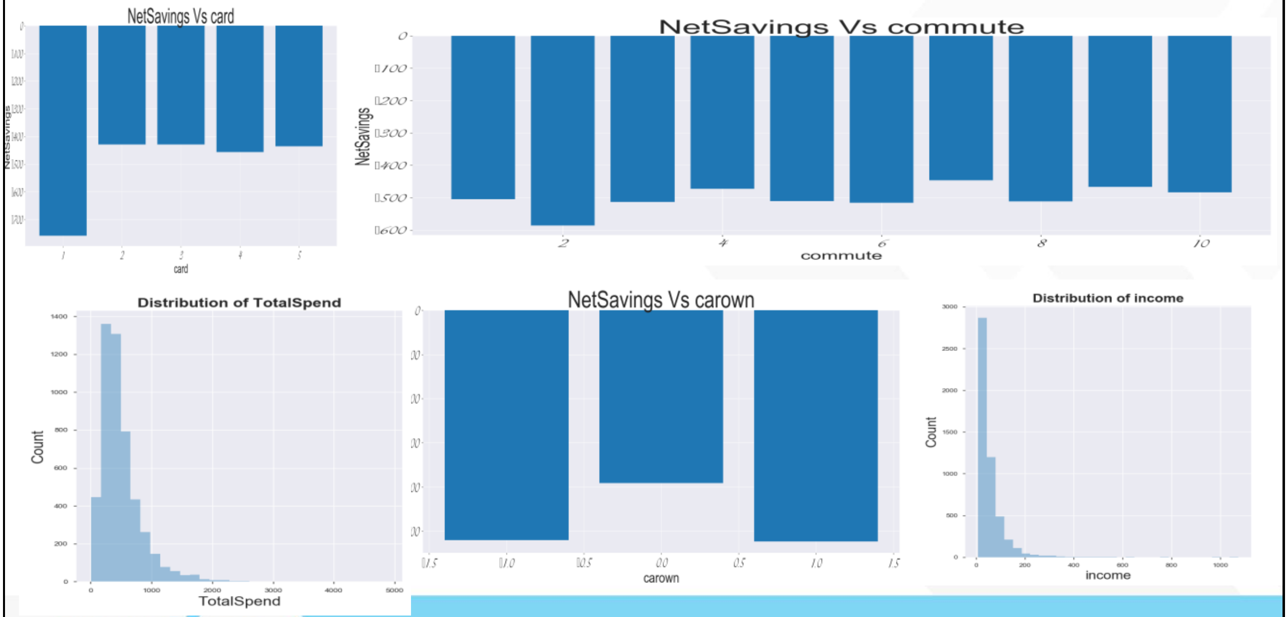
ANALYTIXLABS

Boxplots – TotalSpend & othdebt



Boxplots – Net Savings and income

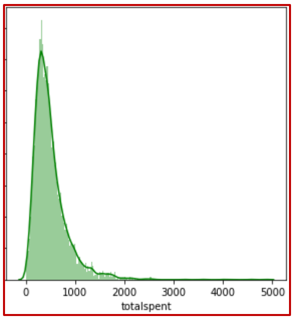
Distributions and variations

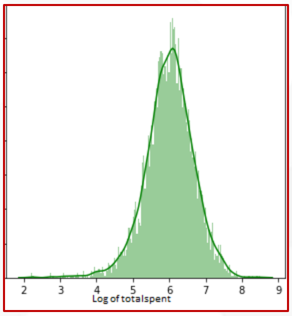


Target(Y) Variable (totalspent)

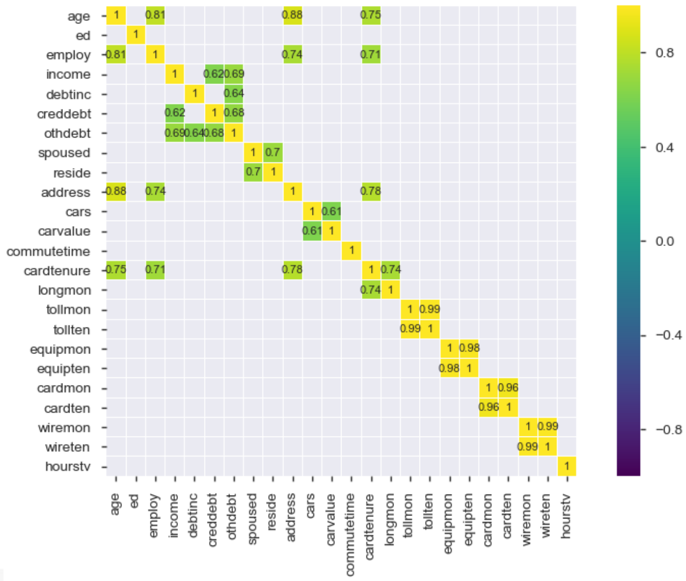
- The Total Spent is skewed right due to some outliers present in the data.
- We will have to convert this distribution to normal for better prediction with machine learning
- We can perform log transformation to convert it to normal distribution

count	5000.00000
mean	498.07863
std	351.52927
min	8.11000
25%	276.28250
50%	414.25000
75%	615.56250
max	4881.05000





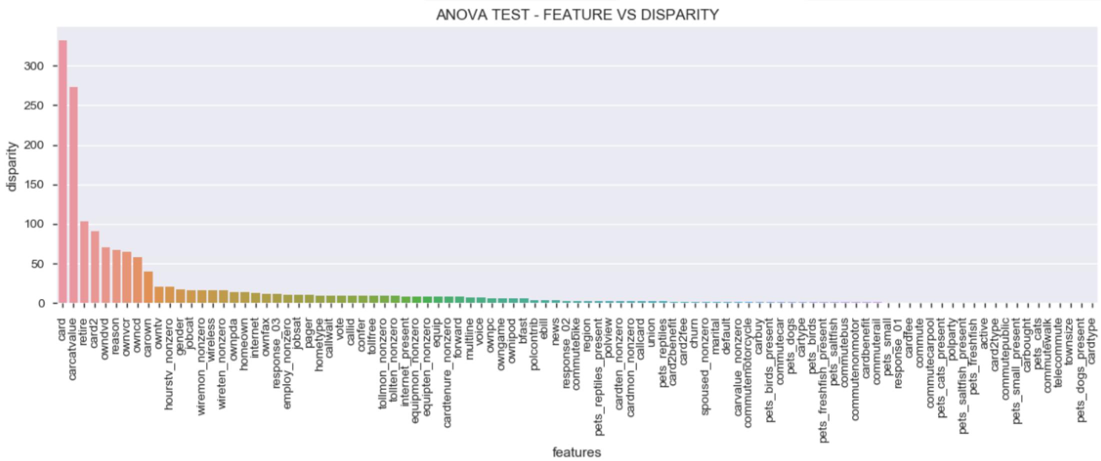
Bivariate Analysis Numerical to Numerical: Co-relation



- Variables income and creddebt and othdebt and carvalue are highly co-related, we can just chose one of them for the model building
- age employ, cardtenure and address are highly co-related, we can choose one of them for the model building
- equipmon and equipten are highly co-related, similar case with cardmon and cardten and wiremon and wireten. One of them can be used for model building

ANALYTIXLABS

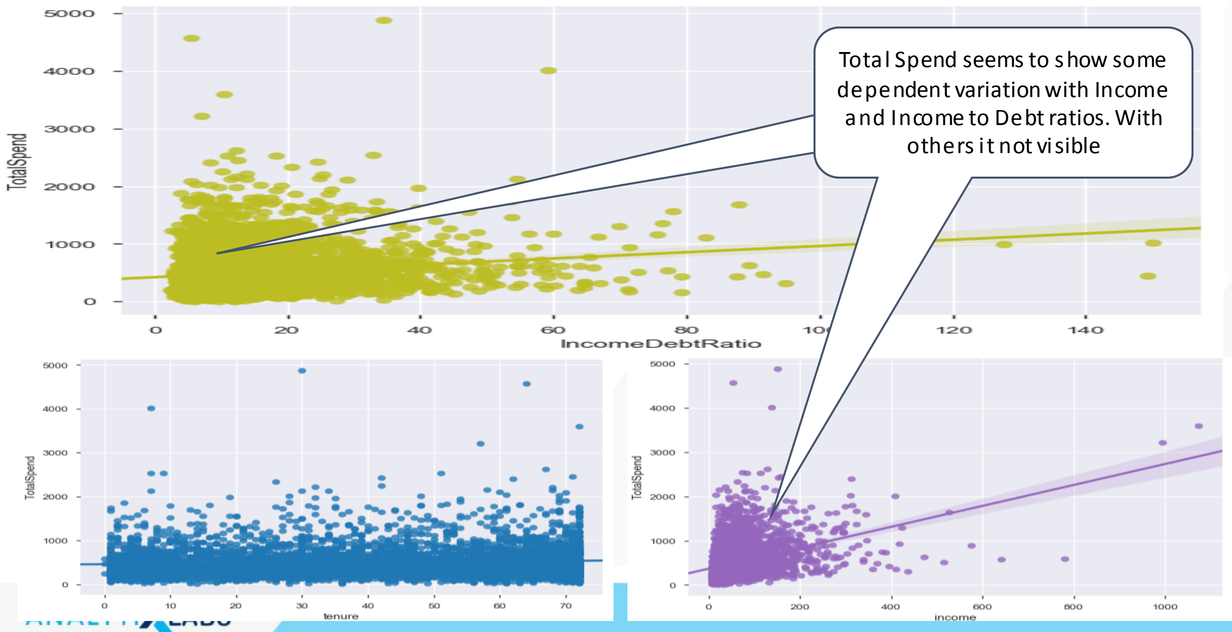
Bivariate Analysis Categorical to Numerical: ANOVA



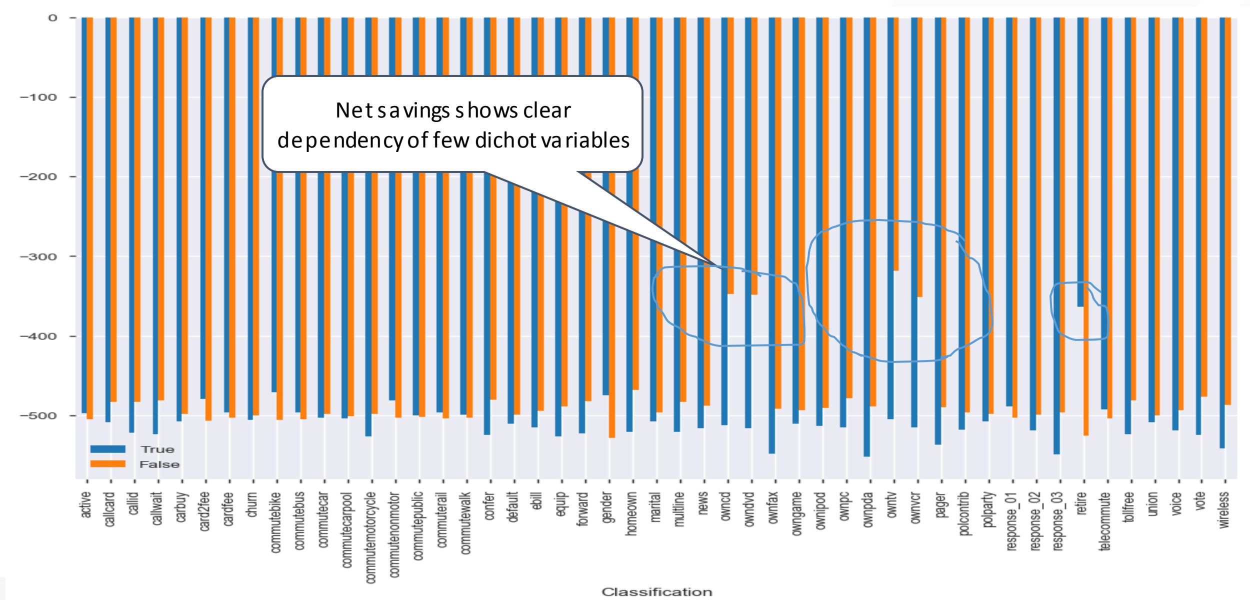
- The ANOVA test shows the relation of the variables with the target variable
- Card seems to be the best variable among all categorical variables

ANALYTIXLABS

Bivariate analysis



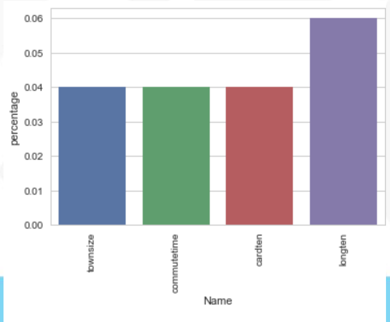
Net Savings vs Dichotomous variables



Data Preparation for Model Building

Data Preparation

- **Target variable:** totalspent = cardspent + card2spent
- **Column Drop:**
    - 2 columns ("custid", "birthmonth") representing personal ID information were dropped
    - All log transformed columns were dropped. Any required transformation will be performed as required during feature engineering steps
  - **Improper Data:**
    - There is variables "carbought" which has -1 as the value. This will be treated as missing value and imputed with mode
    - For spousecat, we can change the value -1 to 6
    - For "carown", "cartype", "carcatvalue", carvalue and spoused, we can change the value -1 to 0
- **Missing Data:**
    - Replacing missing values with median value for continuous variables
    - Replacing missing values with mode value for categorical variables
    - Imputation strategy:
        - Mode: townsize
        - Mean: commutetime
        - Default (0) : longten and cardten



Data Preparation

- **Outlier Treatment:**
  - All the numerical variables were clipped at 1% and 99% percentile values
- **Numerical Transformation:**
  - PowerTransformer (box-cox, yeo-johnson), QuantileTransformer and Log transformation was performed on all numerical variables
  - Best transformation was found as Log transformation
- **Numerical vs Categorical:**
  - Regression check was performed to check the R2 accuracy of each numerical and its categorical against the target variable.
  - Values with better R2 were selected.

Dropped variables:

- 'agecat', 'edcat', 'empcat', 'inccat', 'spousedcat', 'pets', 'addresscat', 'commutecat', 'cardtenurecat', 'card2tenurecat'

```
*************************** R2 ***************************
age : numerical = 4.2542  ,  categorical = 2.7151
ed : numerical = 1.2371  ,  categorical = 0.9483
employ : numerical = 2.3232  ,  categorical = 1.2603
income : numerical = 24.0957  ,  categorical = 14.0609
spoused : numerical = 0.5307  ,  categorical = 0.29
pets : numerical = 0.2574  ,  categorical = 7.5258
address : numerical = 1.9828  ,  categorical = 1.2695
commute : numerical = 0.1733  ,  categorical = 0.0501
cardtenure : numerical = 1.3871  ,  categorical = 0.8421
card2tenure : numerical = 1.4719  ,  categorical = 0.7671
**********************************************************
```

ANALYTIXLABS

Data Preparation – Feature Engineering

- **New Categorical variables:**
  - As there are lot of variables having huge number of zeroes ,we have created few binomial features which represent zero and non-zero values
- **Feature Interactions:**
  - Various feature interactions were created using the PolynomialFeatures function from sklearn.preprocessing.
  - Degree = 2 was used as input.
  - It included Feature to Feature interaction as well as square of variables
- **Polynomial up to Degree 3:**
  - There was separate addition of square and cube versions of the numerical features only in the final data

ANALYTIXLABS

Data Preparation – Feature Engineering

Significant variables based on RFE

- carcatvalue_1
- carcatvalue_2
- carcatvalue_3
- lninc
- inccat
- card2_2
- carown_0
- reason_2
- edcat
- owndvd
- gender
- reason_2
- wireless
- ebill
- marital

Significant variables based on F_regression

- lninc
- inccat
- carcatvalue_1
- carcatvalue_3
- owndvd
- carown_0
- reason_2
- carown_1
- card_2
- card_3
- edcat
- carcatvalue_2
- card2_3
- owntv
- gender

Top 15 variables in each selection method

ANALYTIXLABS

Data Preparation – Feature Engineering

VIF for some of the features

VIF_Factor	features
1.021509838	pets_saltfish
1.02238452	card2fee
1.023634203	cardfee
1.025951369	pets_cats
1.027038075	pets_birds
1.028094475	pets_freshfish
1.029681848	pets_small
1.030027304	pets_reptiles
1.030239551	pets_dogs
1.034243815	response_02
1.040096324	gender
1.045757701	union
1.047804009	response_03
1.050093382	response_01
1.072157909	polcontrib
1.073137107	vote
1.093947876	cartype_0

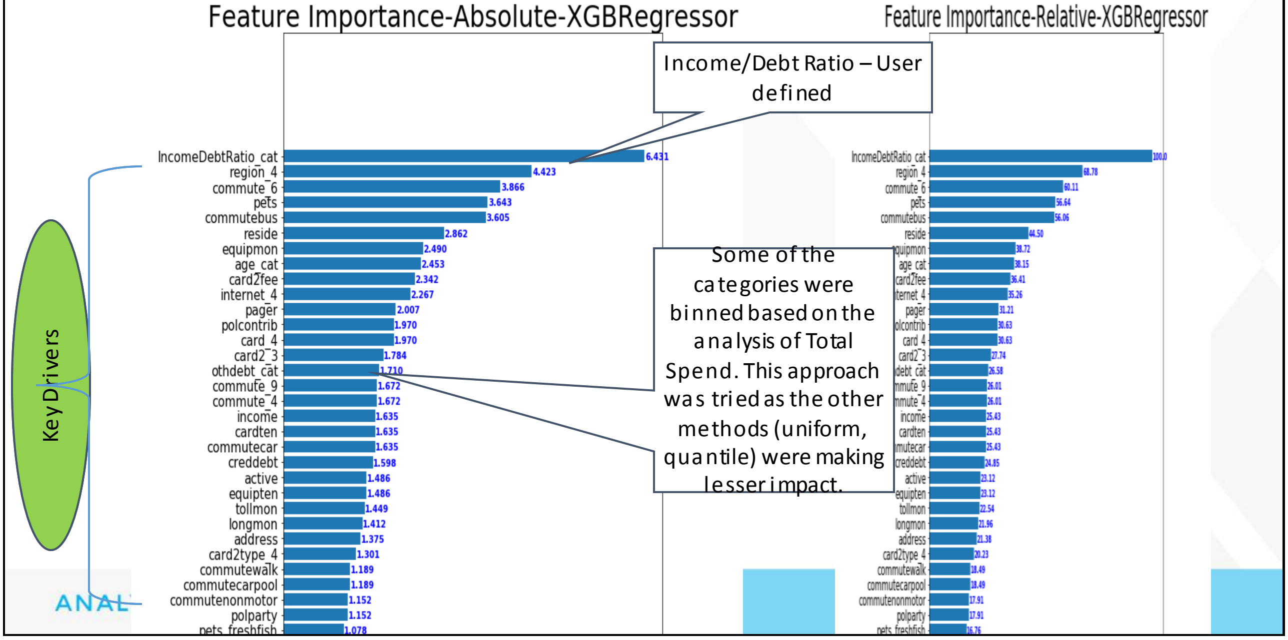
PCA
For some features

Features	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
IncomeDebtRatio	-0.00804	0.009564	-0.09241	-0.01186	0.002859	-0.27889	0.011034	0.056639	-0.07366	0.027141	-0.01037	0.033336	-0.03949	-0.03547	-0.05337	-0.02882	-0.05893
active	0.177897	0.142708	0.14909	-0.03914	-0.13088	0.005539	**-0.71401**	-0.18206	-0.1919	0.233973	0.007525	0.055761	0.019651	0.029847	0.032472	-0.03681	0.172005
address	-0.27666	-0.17697	-0.28338	0.049115	0.016233	**0.700875**	0.115228	-0.21371	0.115948	-0.09476	0.049761	-0.06762	-0.02903	-0.07013	0.038077	-0.03883	-0.05327
address_cat	0.212248	0.198999	0.405956	0.026255	-0.02342	**0.801746**	-0.06862	-0.14734	0.039632	-0.05691	0.012953	-0.03087	-0.03115	-0.01859	0.028061	-0.03266	-0.018
age	-0.20451	0.075097	0.032062	-0.04273	-0.06387	0.009529	-0.004	0.020129	0.063974	-0.10498	0.061649	-0.063	0.00183	-0.05844	0.01536	-0.0106	0.093977
age_cat	**0.759992**	-0.26844	-0.10072	0.147558	0.098825	0.008092	-0.00464	-0.01362	-0.06982	0.149789	-0.06555	0.0763	0.070972	-0.04362	-0.05662	-0.00184	-0.13192
bfast_2	**0.777571**	-0.25312	-0.06071	0.147312	0.105132	-0.00222	-0.01825	-0.01047	-0.08427	0.162636	-0.07468	0.09034	0.071319	-0.02916	-0.05575	0.007876	-0.14623
bfast_3	**0.731229**	-0.34054	-0.14966	0.127369	0.22625	0.017836	0.018051	-0.02349	-0.12283	0.195146	-0.07709	0.08928	0.061166	0.09199	-0.03118	0.030395	-0.13949
callcard	**0.722717**	-0.28041	-0.04401	0.09394	0.208515	0.010888	-0.00543	-0.01631	-0.11895	0.204565	-0.08371	0.09205	0.055817	0.098662	-0.04062	0.028506	-0.1367
callid	**0.528089**	0.081819	-0.24635	-0.24018	-0.06851	-0.0109	0.03235	-0.02268	0.08875	-0.11917	0.025497	-0.06952	-0.02101	-0.05748	0.049579	-0.05733	0.055979
callwait	0.202357	0.428201	-0.1936	**-0.58602**	0.010794	0.027985	0.021253	0.00491	-0.01495	0.019632	-0.00424	0.016312	0.002245	0.00913	0.031777	-0.02027	-0.01066
carbought_0	0.222134	0.417127	-0.17056	**-0.58803**	0.019261	0.019474	-0.00485	0.017931	-0.02029	0.023455	-0.01364	0.016178	0.007818	0.025071	0.008138	-0.0299	-0.03338
carbought_1	0.016004	0.010955	-0.01882	-0.0102	0.015171	0.0232	0.044468	-0.0527	0.144787	-0.03929	-0.35082	0.143101	0.215014	0.031912	0.032489	0.124518	0.130265
carbuy	-0.00127	-0.0051	-0.01781	-0.00926	0.034471	-0.02359	0.028896	0.022652	-0.05797	-0.00687	0.00071	-0.00111	-0.04189	0.000237	-0.0682	0.038623	0.045288
card2_2	0.172858	0.151882	0.359557	0.024706	-0.01769	**0.753098**	-0.05974	-0.13365	0.067346	-0.06554	0.018162	-0.0352	-0.00345	-0.00064	0.048796	-0.01754	0.020173
card2_3	**0.871299**	-0.25852	-0.16183	0.156071	-0.06451	0.000129	0.000547	-0.01077	0.05576	-0.06925	0.032767	-0.04571	-0.01685	-0.03768	0.024348	-0.02494	0.026053
card2_4	0.000423	-0.00777	0.006491	-0.02046	0.002323	-0.01188	-0.0167	0.005327	-0.0247	-0.00582	-0.00564	-0.03414	0.033752	0.073562	0.019849	0.075838	0.029788
card2_5	**0.553716**	0.054052	-0.25027	-0.18865	-0.07597	-0.0108	0.043975	-0.04212	0.129552	-0.14169	0.031501	-0.10404	-0.04201	-0.05319	0.079459	-0.05815	0.132609
card2benefit_2	0.20074	0.19391	0.373844	0.023787	-0.0223	**0.703134**	-0.06185	-0.13124	0.017396	-0.04189	0.007351	-0.0238	-0.04406	-0.02872	0.011321	-0.03721	-0.03798
card2benefit_3	**0.722898**	-0.12823	-0.23861	0.00443	-0.10138	-0.00309	0.035893	-0.03678	0.140373	-0.15455	0.054412	-0.11851	-0.04759	-0.06622	0.077587	-0.06367	0.140136
card2benefit_4	-0.45667	0.095279	0.049898	-0.09538	0.12274	0.02837	0.03901	-0.02978	-0.03914	0.130982	-0.02443	0.061487	-0.00707	-0.0292	-0.02795	-0.01003	0.016342
card2fee	**0.87707**	-0.27468	-0.16069	0.162135	-0.03656	0.003462	-0.00144	-0.00777	0.037845	-0.03922	0.017062	-0.03316	-0.01042	-0.03246	0.0261	-0.01754	0.015922
card2tenure	**0.851842**	-0.23853	-0.11996	0.171679	-0.04822	-0.01622	-0.03924	0.01522	0.017968	-0.05024	-0.00399	-0.01949	-0.00242	-0.00649	0.022108	-0.00187	-0.02246
card2type_2	0.036861	0.039538	0.188873	-0.01638	0.033312	-0.26692	0.112872	**-0.71106**	0.219017	-0.02334	0.065006	-0.13039	-0.04531	-0.03405	-0.04037	-0.0297	-0.16047
card2type_3	0.495545	0.302706	**0.643094**	-0.02923	0.043682	-0.11778	-0.0275	-0.17259	-0.01588	0.180303	0.020791	0.029253	0.003006	0.03545	-0.00141	-0.00202	0.088987
card2type_4	0.464355	0.310962	**0.638758**	-0.02482	0.017567	-0.14016	-0.08391	-0.17727	-0.01175	0.133122	0.005335	0.030456	0.004988	0.03773	-0.00482	0.010965	0.057139
card_2	-0.31885	0.352842	0.004652	0.143632	0.097044	0.002482	0.022404	-0.08086	-0.04605	0.088256	-0.04753	0.03792	-0.00753	0.047549	-0.01149	-0.0079	0.005696
card_3	-0.0203	0.022582	-0.01331	0.013343	-0.0314	-0.04821	-0.02534	0.067882	0.207552	0.14593	-0.0008	-0.03502	0.011438	0.014077	0.027326	0.001817	-0.01779

ANALYTIXLABS

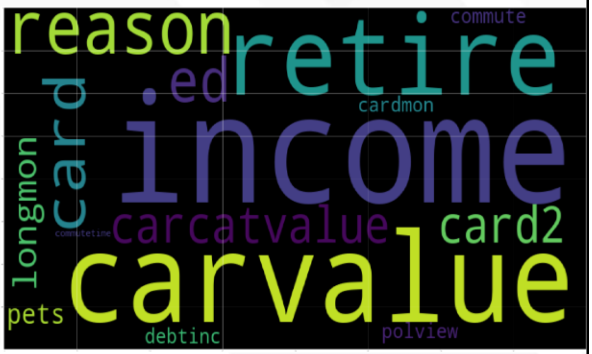
Feature importance

(Based on final model- XG Boost)



Data Preparation - Feature Reduction

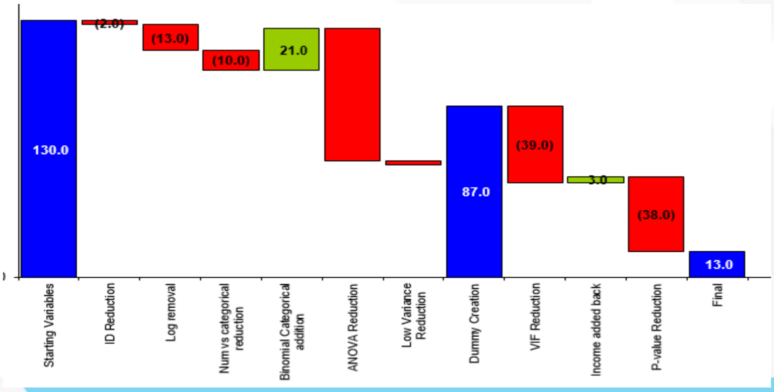
- We started with all features that are output of the data preparation steps.
- Finally we found out the best features by using three methods **Factor analysis, RFE** and **Random Forest**
- Multicollinearity was addressed using VIF with threshold = 5. As an business exception income was added back in the final features list



ANALYTIXLABS

Feature Reduction - Waterflow chart for variable selection

- The variables are added and reduced in various stages as shown below.
- The final step of P-value reduction is shown for LinearRegression.
- In case of RandomForest and XGBoost it is performed using feature importance
- **Note**: Feature interaction addition and RandomForest Ranking reduction are not shown below for better legibility. Around 2200 features were added and then reduced as mentioned above.



Model Building & Fine Tuning

Model Building Approaches

- Below models were explored
  - Linear Regression
  - Random Forest Regression
  - Adaboost Regression
  - Gradient Boost Regression
  - XG Boost Regression
- Other approaches
  - Tried predicting primary and secondary card spends separately and then adding up
  - Many variations of y variable were tried such as Spend ratio (Total Spend/Income), Spend Percentage, Spend difference. All had minor variations but not significant.
  - Calculating median spend based on certain categories. But this was dropped as this was not helpful in deciding drivers. Although this was giving improved R2 and reduced MAPE
  - Combining XG Boost and then successive application of Adaboost and Gradient boost gave improved MAPE
  - Transformation techniques – StandardScaler & PowerTransform. However, both had limited impact

Data Split and Model Executions

- Data was split into Train(70%) and Test(30%).
- There was k-Fold validation performed where k-fold was taken as 10
- Baseline was defined by taking mean value of train data as prediction
- Three techniques **LinearRegression**, **RandomForestRegressor** and **XGBRegressor** were used to check the MAPE and Accuracy.
- Detail execution and corresponding outputs are present in attached document.

Credit Card Spend model executions

Models Comparison - Metrics

- XGBClassifier model has come out as the top model with consideration of train R2, MAPE_ACCURACY.
- LinearRegression stands second with small reductions in accuracies but better in terms of reduced overfitting

Final Model:

- As there is not much difference in the outcomes of Linear Regression and XGBoostClassifier models, either of them can be chosen as final model.
    - But due to less overfitting, we will finalize Linear Regression as the best model

Best parameters after tuning:

- ***Linear Regression*** *:'copy_X': True, 'fit_intercept': True, 'normalize': False*
- ***XGBoost*** *: learning_rate=0.01, n_estimators=686, max_depth=4, min_child_weight=1, gamma=1, subsample=0.8, colsample_bytree=1, reg_alpa=1.2, scale_pos_weight=1*
- ***RandomForest****: n_estimators=240, bootstrap=True, max_features='auto', max_depth=8 min_samples_split=50, min_samples_leaf=7,*

Model Name	Features_used	Metric	Test	Train
BASELINE	ORIG	MAE	235.7	237.35
		MAPE_ACCURACY	24.98	23.52
		R2_ACCURACY	NA	0.00
LinearRegression	FINAL	MAE	188.8	186.28
		MAPE_ACCURACY	52.62	53.09
		R2_ACCURACY	NA	32.10
	ORIG	MAE	189.19	185.90
		MAPE_ACCURACY	52.5	53.31
		R2_ACCURACY	NA	32.67
RandomForestRegressor	FINAL	MAE	195.75	177.76
		MAPE_ACCURACY	50.04	55.84
		R2_ACCURACY	NA	34.72
	ORIG	MAE	198.98	78.69
		MAPE_ACCURACY	49.96	83.46
		R2_ACCURACY	NA	84.57
XGBRegressor	FINAL	MAE	191.46	175.02
		MAPE_ACCURACY	52.06	57.34
		R2_ACCURACY	NA	36.87
	ORIG	MAE	195.74	137.97
		MAPE_ACCURACY	51.18	67.55
		R2_ACCURACY	NA	59.74

ANALYTIXLABS

Model Comparison – Other models

Final Model

Model Comparison													
				Train					Test				
Model Group	Features	Model	Standardization	R^2	MAE	MSE	RMSE	MAPE	MAE	MSE	RMSE	MAPE
1	PCA	Linear Regression	StandardScaler	0.2414	209.07	82060	286	62.1	211.1	80993	285	64.36
2	PCA	XG Boost	StandardScaler	86.7	87.8	14286	119.5	27.1	214.4	83464	289	65.62
3	All	Gradient Boost	StandardScaler	45.8	179	58631	242	53	194.6	69469	264	58.8
3	All	Layered Adaboost	StandardScaler	45.9	175.4	59535	244	47.7	190.6	69832	264	53.6
3	All	Layered Adaboost	StandardScaler	43.11	175.5	61535	248	44.7	189.6	71388	267	50.3
4	All	XG Boost	StandardScaler	56.41	160.4	47152	217	48.41	198	70886	266	60.2
4	All	Layered Adaboost	StandardScaler	56.6	156.8	47901	218.9	43.1	193.6	71065	266	55
4	All	Layered GradientB	StandardScaler	54.1	157	49637	223	40.3	192.6	72440	269	51.9
5	All	XG Boost	None	56.1	159.6	47498	218	47.7	196	70257	265	59
5	All	Layered Adaboost	None	55.2	157	48479	221	42.8	192	70723	266	54
5	All	Layered GradientB	None	53.6	157	50245	224	40.2	191	72170	268	51.1

ANALYTIXLABS

Tool of implementation & Key Drivers

- The implementation tool is created in excel and attached below.
- Final equation using Linear Regression:
  - 5.1113 - 0.5884 * card_2 - 0.6011 * card_3 - 0.7023 * card_4
    - 0.4905 * card_5 - 0.3913 * card2_2 - 0.3831 * card2_3
    - 0.4361 * card2_4 - 0.2813 * card2_5 + 0.2483 * reason_2
    - 0.165 * reason_4 - 0.0553 * gender_1 + 0.5838 * income
    - 0.0305 * income_sqr

Implementation Tool

- Credit Limit will be given thrice of predicted Credit card spend.
- Income is the major positive key driver for finding credit limit of any customer
- Having American Express card increases the credit limit by around 2-3 times compared to other cards.
- Male customers are given more credit compared to female customers
- Customers coming to get new credit cards with reason of "Convenience" get better credit limit
- Income _square helps to put a cap on credit limit for high income customers
- All customers irrespective of deciding factors will get minimum credit limit of 165

Positive drivers are marked in green
Negative drivers are marked in red

ANALYTIXLABS

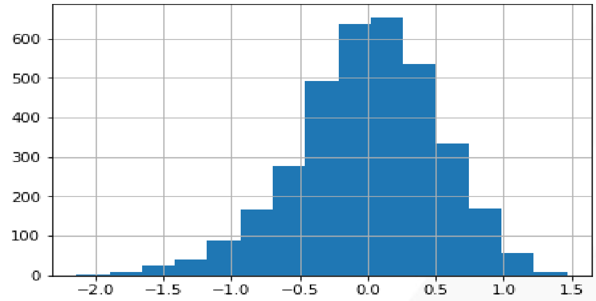
Decile Analysis

Decile	pred_val	Cardspent_val
9	833.5420	912.2399
8	621.8103	712.9803
7	529.5894	593.1252
6	464.9185	517.2574
5	413.2706	465.7444
4	372.2152	425.1774
3	336.5317	388.5363
2	307.3986	343.4760
1	275.1145	319.2682
0	223.0948	263.8456

Decile	Pred_val	Cardspent_val
9	860.0851	915.7990
8	628.8650	707.0851
7	530.9035	538.8130
6	464.4957	536.6252
5	415.6433	483.7722
4	373.9221	423.9477
3	338.4644	374.7855
2	304.5084	344.6931
1	274.5320	317.2925
0	226.7048	274.5087

ANALYTIXLABS

Residual Plots

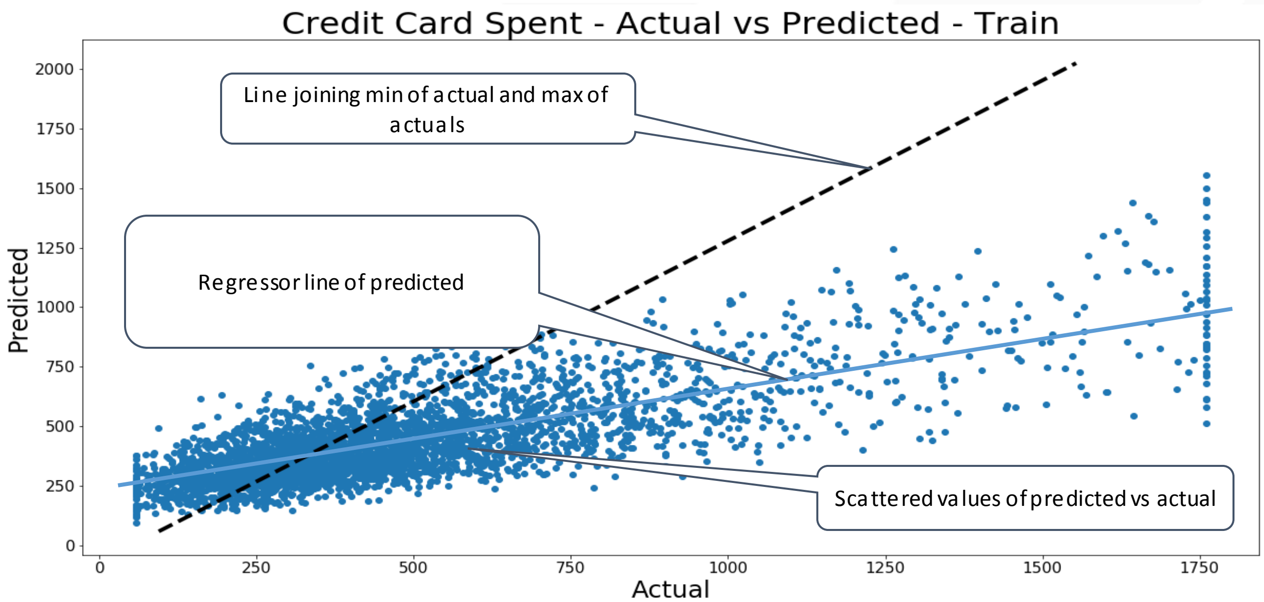


From residual plot we observe that the residuals are follows normal distribution.

Model is working well on both training and testing data sets and residuals are follows normal distribution.

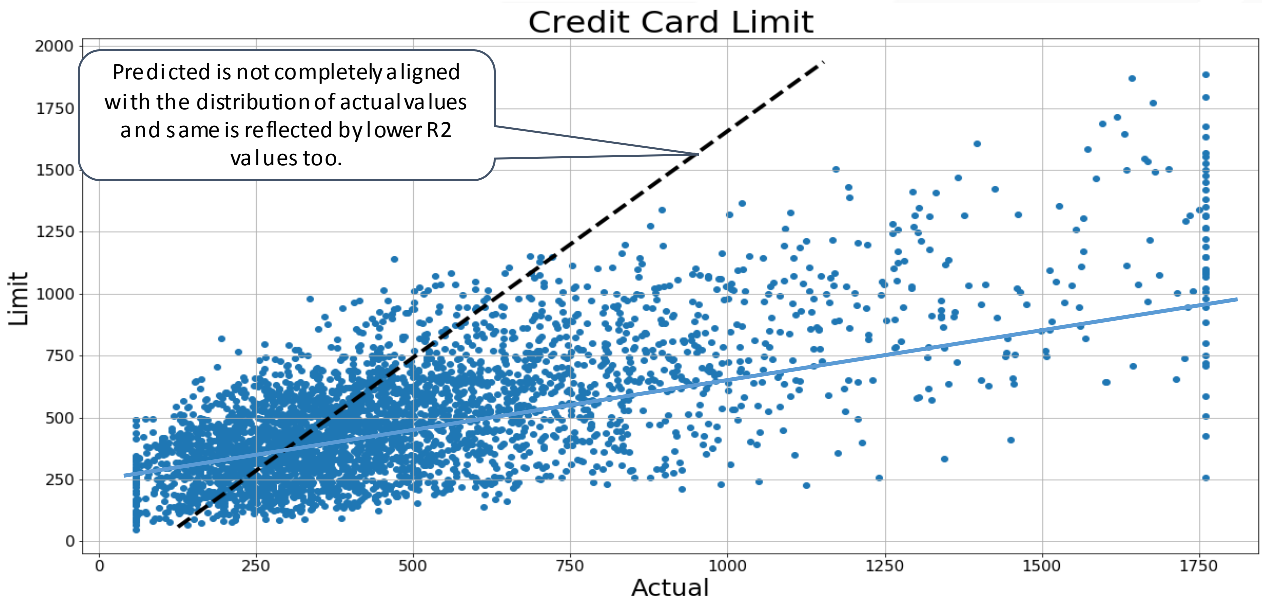
ANALYTIXLABS

Model Prediction - Train Data

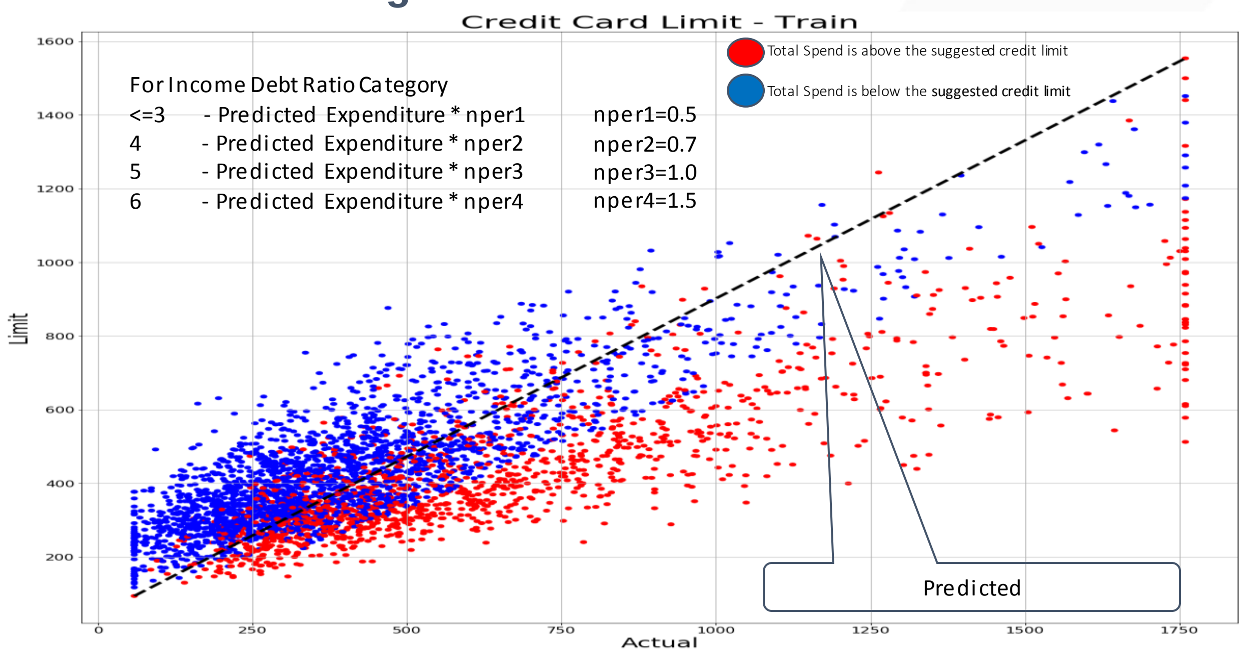


ANALYTIXLABS

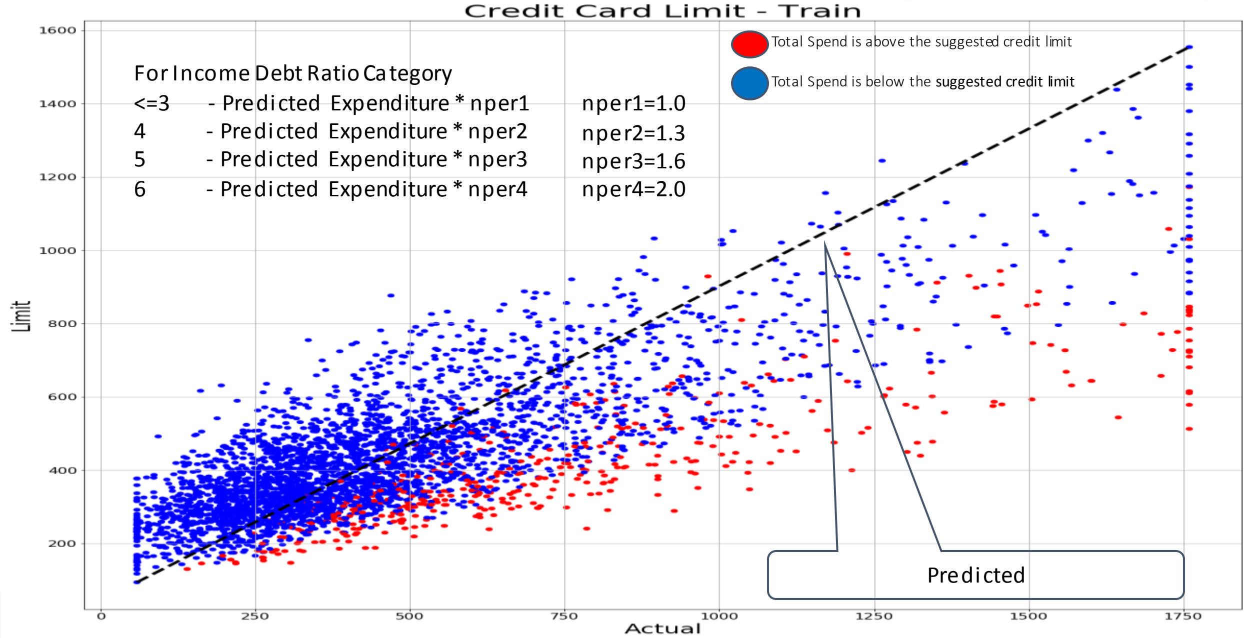
Model Prediction – Test Data



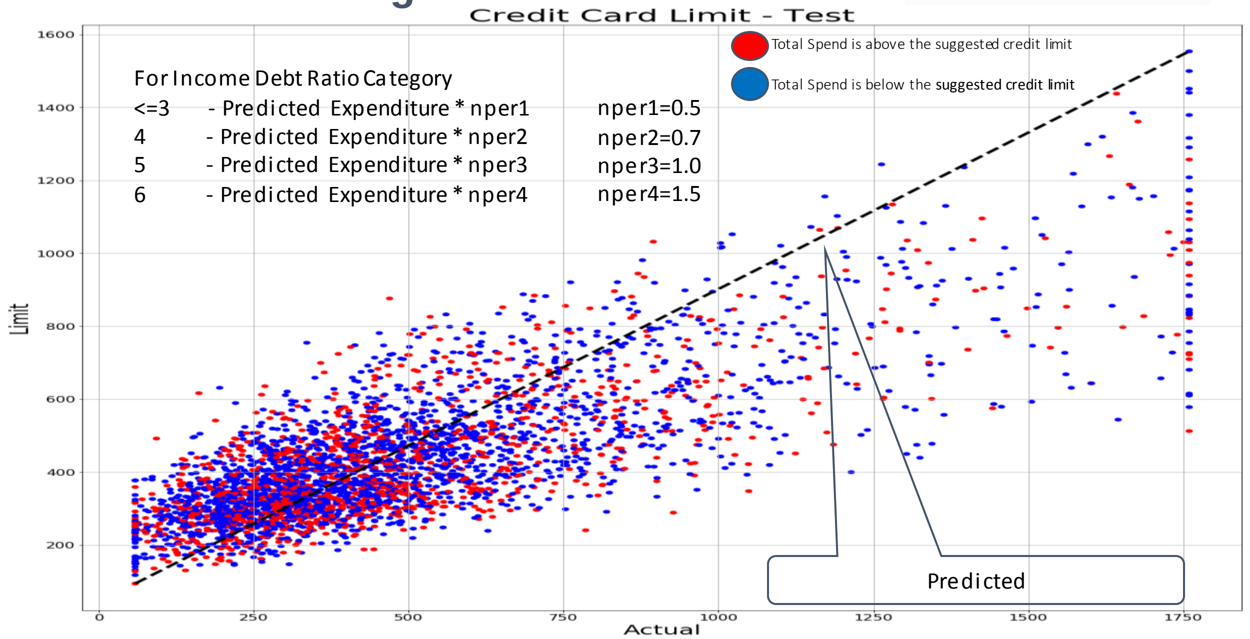
Credit Limit – Stringent limits - Train



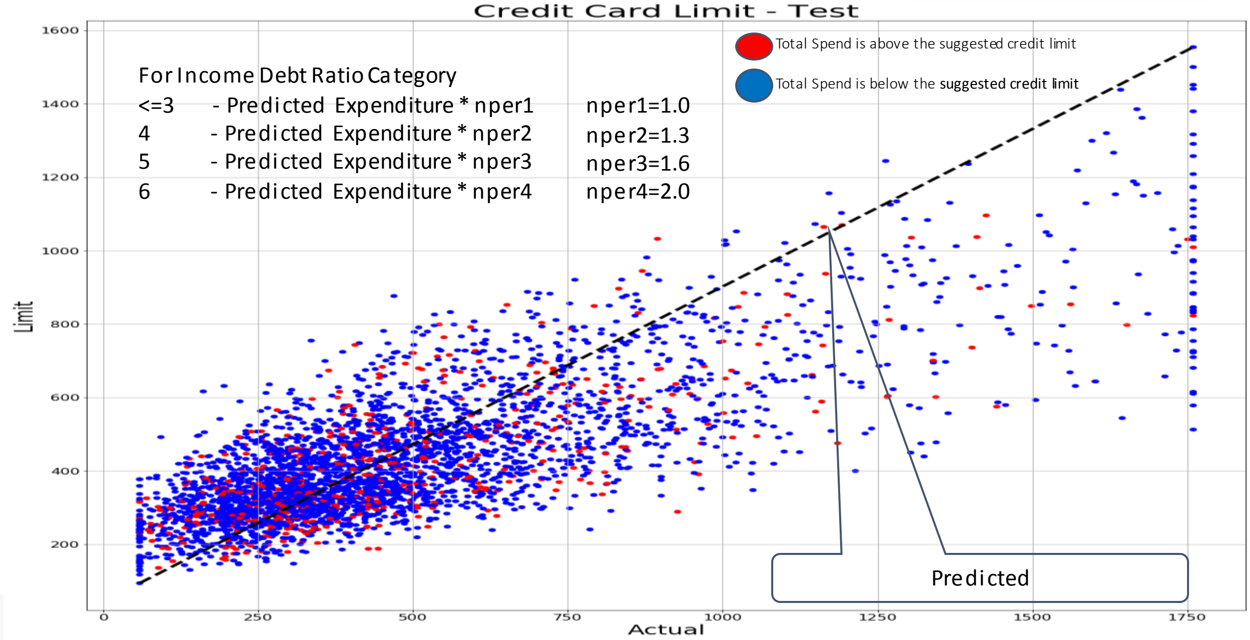
Credit Limit – Lenient limits - Train



Credit Limit – Stringent limits - Test



Credit Limit – Lenient limits - Test



Credit Card Spend prediction and assign Credit Limit

Pre requisites & Assumptions :

Data has to be in the prescribed format in excel files.
Missing values would be imputed based on original dataset
XGBoost model pickle would be used for prediction.

Few Configurations

```
In [10]: fname="Data_Set_Test.xlsx"
         fname_old="Data Set.xlsx"  #Required for missing value imputation in case data is missing in the provided data set.
         fpath="E:\\DSA\\13_Assessments\\Project\\CreditCardSpend\\"
         fpath_pickle="E:\\DSA\\13_Assessments\\Project\\CreditCardSpend\\Outputs\\"
         fname_xgb2          = 'bfast_xgb_grid2_20190112-1900PM.pkl'
         fname_add_columns   = 'bfast_create_additional_columns_model_20190112-1902PM.pkl'
```

Functions used for some of the preprocessing

Missing value imputation is based on the original dataset

Predict Credit Card Spend and Credit Limits

```
In [28]: xgb2       =openpickle(fpath_pickle,fname_xgb2)
```

```
In [44]: #Below Parameters can be configured
         nper1=0.5
         nper2=0.7
         nper3=1.0
         nper4=1.5

         y_pred_actual='predict'
         test['predict']=xgb2.predict(test_X)
         ylimit_test=np.where(test['IncomeDebtRatio_cat']<=3,test[y_pred_actual]*nper1,
                        (np.where(test['IncomeDebtRatio_cat']==4,test[y_pred_actual]*nper2,
                           np.where(test['IncomeDebtRatio_cat']==5,test[y_pred_actual]*nper3,test[y_pred_actual]*nper4))))
         test['limit']=ylimit_test
```

```
In [45]: test[['predict','limit']]
```

Out[45]:

	predict	limit
0	439.915436	439.915436
1	412.525024	288.767517
2	574.272400	574.272400
3	305.617828	305.617828
4	347.725891	521.588867
5	473.724335	710.586487
6	559.982727	839.974121
7	622.325439	933.488159
8	330.874329	496.311493
9	575.156006	862.734009
10	378.053558	567.080322
11	308.005157	462.007751
12	546.592285	819.888428
13	579.177917	868.766846
14	314.173248	471.259888
15	376.781616	565.172424
16	633.617493	950.426270

ANALYTIX LABS

Recommendations & Next Steps

- Regular model rebuilding after specific time interval will be required for maintenance.
- The key drivers for the model are income and existing cards possessed by the customer.
- One should make sure that they remain consistent after certain time period. Any changes should be incorporated into the model and retrained accordingly.
- Also check should be performed if newer predictors can be found and helpful to have better prediction on the data.

ANALYTIXLABS

ANALYTIXLABS