

Hints for Final Project: Predicting Credit Card Spend

Disclaimer: This material is protected under copyright act AnalytixLabs ©, 2011-2019. Unauthorized use and/ or duplication of this material or any part of this material including data, in any form without explicit and written permission from AnalytixLabs is strictly prohibited. Any violation of this copyright will attract legal actions

Hints: Predicting Credit Card Spend

1. Create a new variable (Total Spend = Card1Spend (primary card) + Card2Spend(secondary card)), drop card1spent and card2spend). Along with, you can also drop variables like card1items, card2items (because these variables are directly proportional to spend). Apply log transformation for Total spend if it is not following normal distribution. Then your new dependent variable is log_Total_Spend and others are independent variables (you should not include total spend as part of independent or dependent variables when you build the model)
2. Identify categorical & numerical variables from the data as per the data dictionary. Remove duplicate features like log variables, categorical features which are having numerical features etc. Perform data preparation steps on the data (missing treatment, outlier treatment, variable dropping based on low variance etc.)
3. Apply correlation metrics (between Y & X) or factor analysis for reducing numerical variables. Apply anova (between Y and each categorical variable) and identify important categorical variables based on F-value. Apply other variable reduction techniques like F-regression, RFE, KBest Techniques. Consolidate all the variables selected from above methods. Apply VIF for the variables selected, remove the variables based on high VIF in one by one.
4. Apply transformation on numeric Variables (log, sqrt, square, exp) and identify which transformation gives you best correlation between dependent and independent variables.
for example: Income is independent variable, create transformed variables log_income, square_income, sqrt_income, exp_income etc. then calculate correlation between dependent and each of these transformed variables, identify which variables gives you best correlation. That will be your best transformation. incase of log transformation, you can transform the data using $\log(x+1)$ instead of $\log(x)$ to avoid the problem of zero's.

Hints: Predicting Credit Card Spend

5. Split the data into train & test (70:30 pct) Build the model on the development data using above important variables from categorical (after creating dummies) from step 3 and numerical variables from step 4. As a best practice, You should restrict number of variables as less than 15 variables. Once you are sure with all the variable significance, no multi-collinearity, good r-square (in this case ~40%), then proceed with next step (validation). You can build the models using different modeling techniques like DT, RF, GBM, XGBOOST, SVR, KNN etc. Tune the hyper parameters & perform cross validation to avoid over fitting and under fitting problems.
6. Using the final model equation from step 5, score the values (predict the values) for development & validation data sets. Check MAPE, RMSE for development & validation data sets. Also, you can calculate the correlation between actual & predicted values for both development and validation data sets. You can also Compare decile analysis between development, validation in terms of rank ordering and similar values for similar decile's etc.
7. Importance of factors can be calculated using T-value from model results. This will help you to sort the variables based on the contribution. Also you can use feature importance comes from python code to understand the importance of variables.
8. Share all the results in workbook along with outputs & presentation

Contact us

Visit us on: <http://www.analytixlabs.in/>

For course registration, please visit: <http://www.analytixlabs.co.in/course-registration/>

For more information, please contact us: <http://www.analytixlabs.co.in/contact-us/>

Or email: info@analytixlabs.co.in

Call us we would love to speak with you: (+91) 88021-73069

Join us on:

Twitter - <http://twitter.com/#!/AnalytixLabs>

Facebook - <http://www.facebook.com/analytixlabs>

LinkedIn - <http://www.linkedin.com/in/analytixlabs>

Blog - <http://www.analytixlabs.co.in/category/blog/>