Machine Learning Engineer Nanodegree
Capstone Proposal
Brian Cleary
Nov 1, 2015

Domain Background:

Banco Santander (Santander Bank) is currently interested in improving their personalized product recommendations for their banking customers—Kaggle Competition. Santander Group (parent company) servers more than 100 million customers in the UK, Latin America and Europe. According the Banco Santander's website, "we aim to make your banking life easier by providing convenient and smart ways to spend, save, and manage your money – from basic checking and savings accounts to comprehensive financial solutions."

According to the a research report from Aberdeen Group, financial services companies that used predictive analytic saw a 10% increase in identifying new customer opportunities in 2014, compared to a 7% increase in firms not using predictive analytics (TIBCO.) Banco Santander is attempting to improve product recommendations for existing customers to increase the customer experience and increase cross-sell opportunities. The problem can be solved by using predictive analytic algorithms to train a model on past customer behavior in order to target customers most likely to want/need a new financial product. For example, once the algorithm is trained to identify the most likely customers will open a Home Equity Line of Credit (HELOC) in the future, the marketing department can develop integrated marketing campaign to target these customers. By narrowing down the list of potential customers, the company can increase profit by lowering the amount of postcards or outbound calls to customers who are not likely to need the service. Combined with a facial recognition, bank tellers could even pull up the customer's profile in order to cross-sell the product the customer is most likely to want/need (The Financial Brand).

Works Cited:

TIBCO Software Inc. "Predictive Analytics for Financial Services Firms: Forecasting the Future." *Tibco.com*. Web. 27 Jan. 2014. Available at: http://www.tibco.com/blog/2014/01/27/predictive-analytics-for-financial-services-firms-forecasting-the-future/

The Financial Brand. "10 Branch Banking Innovation Strategies for 2016." thefinancialbrand.com. Web. 24 Nov. 2015Available at: https://thefinancialbrand.com/55561/branch-banking-innovation-strategies/

Problem Statement:

Under their current product recommendation system, a small number of Santander's customers receive many recommendations while many others rarely see any resulting in an uneven customer experience. Santander is challenging Kagglers to predict which products their existing customers will use in the next month based on their past behavior and that of similar customers.

Datasets and Inputs:

Per Kaggle, “Banco Santader has provided 1.5 years of customer behavior data. The data starts at 2015-01-28 and has monthly records of products a customer has, such as "credit card", "savings account", ect. Datasets are public and available at: https://www.kaggle.com/c/santander-product-recommendation/data

Column Name	Description	Column Name	Description
fecha_dato	The table is partitioned for this column	ind_ahor_fin_ult1	Saving Account
ncodpers	Customer code	ind_aval_fin_ult1	Guarantees
ind_empleado	Employee index: A active, B ex employed, F filial, N not employee, P pasive	ind_cco_fin_ult1	Current Accounts
pais_residencia	Customer's Country residence	ind_cder_fin_ult1	Derivada Account
sexo	Customer's sex	ind_cno_fin_ult1	Payroll Account
age	Age	ind_ctju_fin_ult1	Junior Account
fecha_alta	The date in which the customer became as the first holder of a contract in the bank	ind_ctma_fin_ult1	Más particular Account
ind_nuevo	New customer Index. 1 if the customer registered in the last 6 months.	ind_ctop_fin_ult1	particular Account
antiguedad	Customer seniority (in months)	ind_ctpp_fin_ult1	particular Plus Account
indrel	1 (First/Primary), 99 (Primary customer during the month but not at the end of the month)	ind_deco_fin_ult1	Short-term deposits
ult_fec_cli_1t	Last date as primary customer (if he isn't at the end of the month)	ind_deme_fin_ult1	Medium-term deposits
indrel_1mes	Customer type at the beginning of the month ,1 (First/Primary customer), 2 (co-owner),P (Potential),3 (former primary), 4(former co-owner)	ind_dela_fin_ult1	Long-term deposits
tiprel_1mes	Customer relation type at the beginning of the month, A (active), I (inactive), P (former customer),R (Potential)	ind_ecue_fin_ult1	e-account
indresi	Residence index (S (Yes) or N (No) if the residence country is the same than the bank country)	ind_fond_fin_ult1	Funds
indext	Foreigner index (S (Yes) or N (No) if the customer's birth country is different than the bank country)	ind_hip_fin_ult1	Mortgage
conyuemp	Spouse index. 1 if the customer is spouse of an employee	ind_plan_fin_ult1	Pensions
canal_entrada	channel used by the customer to join	ind_pres_fin_ult1	Loans
indfall	Deceased index. N/S	ind_reca_fin_ult1	Taxes
tipodom	Addres type. 1, primary address	ind_tjcr_fin_ult1	Credit Card
cod_prov	Province code (customer's address)	ind_valo_fin_ult1	Securities
nomprov	Province name	ind_viv_fin_ult1	Home Account
ind_actividad_cliente	Activity index (1, active customer; 0, inactive customer)	ind_nomina_ult1	Payroll
renta	Gross income of the household	ind_nom_pens_ult1	Pensions
segmento	segmentation: 01 - VIP, 02 - Individuals 03 - college graduated	ind_recibo_ult1	Direct Debit

Solution Statement:

Per Kaggle, I will predict what additional products a customer will get in the last month, 2016-06-28, in addition to what they already have at 2016-05-28. These products are the columns named: ind_(xyz)_ult1, which are the columns #25 - #48 in the training data. I will predict what a customer will buy in addition to what they already had at 2016-05-28." **Quantifiable**: The problem can be expressed mathematically by using the data provided and converting any categorical data using dummy variables in order to run classification algorithms to predict likelihood any given customer will choose a new product in the next month. I can then use the likelihood for each product to rank-order the products most likely to be chosen. **Measurable:** The metric discussed below is MAP@7 which will be used the train the model on existing data and judge the model's performance on the test set (once I submit predictions to Kaggle). **Replicable**: The problem can be reproduced each month by Santander; as the company collects all relevant data and can make predictions each month based on the model—due to seasonality, it is advised to determine what products have increased demand depending on the month.

Benchmark Model

According to a Michael, a Udacity lecturer, Naïve Bayes is often a good first model to run because it provides a reasonable baseline to compare other models against. I will use the Naïve Bayes classifier as a benchmark because it has several strengths: it is easy to implement, well known and understood, and empirically successful. If independence of attributes holds, NB classifier will converge quicker than discriminative models.

Naïve Bayes biggest disadvantage is its simplicity. At times, model complexity is advantageous when generalizing the relationship of from complex relationships and interactions. Furthermore, the assumption of independence among attributes may not be realistic and models that do not force this assumption on the class structure can better perform when there is high dependence among attributes. According to Haste, even if the individual class density estimates are biased, the posterior probabilities near the decision boundary can withstand considerable bias—the posterior probabilities can be smooth even when the population class densities are not (pg. 210-211).

Works Cited:

H. Liu, X. Yin and J. Han, "An Efficient Multirelational Naive Bayesian Classifier Based on Semantic Relationship Graph", ACM MRDM, 2005.

Hastie, Trevor, Trevor Hastie, Robert Tibshirani, and J H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer, 2001. Print.

Evaluation Metrics

Per the **Kaggle** completion, I will use the MAP@7 metric—MAP is just a mean of average precision for all users. In other words, if we have 1000 users, we sum APs for each user and divide the sum by 1000. Each Banco Santander customer is hypothetically interested in some "new" products. We are tasked with recommending 7 items per user. In this completion, MAP@7 indicates the MAP for up to 7 product recommendations per customer. I am not penalized for bad guesses, so submitting all 7 recommendation is preferred; however, order matters (unless I get all right). I will select the best 7 candidates per customer (in order of most likely products to least likely).

Equations for AP and MAP: Average precision at n for the user**—**P (k) means the precision at cut-off k in the item list, i.e., the ratio of number of users followed, up to the position k, over the number k; P(k) equals 0 when the k-th item is not followed upon recommendation; m is the number of relevant nodes; n is the number of predicted nodes. If the denominator is zero, P(k)/min(m,n) is set to zero. The mean average precision is the average of the AP at n for each user.

$$ap@n = \sum_{k=1}^{n} P(k)/min(m, n)$$

$$MAP@n = \sum_{i=1}^{N} ap@n_i/N$$

Works Cited:

Kaggle. “Mean Average Precision” *Kaggle.com*. Web. N.d. Available at: https://www.kaggle.com/wiki/MeanAveragePrecision

Kaggle. “Santander Product Recommendation” Kaggle.com. Web. 26 Oct. 2016. Available at: https://www.kaggle.com/c/santander-product-recommendation/details/evaluation

Z. Zygmunt. “What you wanted to know about Mean Average Precisionn” fastml.com. Web. 08 Aug. 2012. Available at: http://fastml.com/what-you-wanted-to-know-about-mean-average-precision/

Project Design

Theoretical Workflow per Kaggle’s post, *How to get into the top 15 of a Kaggle completion using Python:*

1. Download datasets from Kaggle (train_ver2.csv, test_ver2.csv, sample_submission.csv)
2. Review the column description
3. Explore Santander website to better understand the products offered
4. Figure out what I am predicting: I will predict what additional products a customer will get in the last month, 2016-06-28, in addition to what they already have at 2016-05-2
5. Create/Use a scoring algorithm script for MAP@7
6. Use visualizations and cross tabs for EDA:
    a. What products are most used?
    b. Is there seasonality?
    c. Do people add/drop products often?
    d. Flag deceased customers
7. Confirm all test-set customer code (customer Id) are in training set
8. Determine if I can down sample the Kaggle Data in order experiment with different techniques
    a. Ideally want small enough dataset to quickly iterate different approaches
    b. Random sampling of rows (train)—to create training, cross validation (if needed), testing sets.
9. Pick 10,000 users: “Because the customer_code in test are a subset of the customer_code in train, we’ll need to do our random sampling in a way that preserves the full data of each user. We can accomplish this by selecting a certain number of users randomly, then only picking rows from train where customer_code is in our random sample of user ids.
10. Generate predictions for benchmark – NB Classifier
11. Score benchmark with MAP@7 metric
12. Explore correlations for products and other variables
    a. Rules out logistic regression unless linear correlations high (unlikely)

13. Determine if dimension reduction techniques (PCA) will be useful in simplifying the feature space
14. Feature engineering:
    a. Potential use KMean clustering on PCA to generate a new feature that can be used in the training and test sets
    b. Perform domain knowledge research to determine if any additional features can be found and included (unsure what those are right now).
15. Use cross validation and grid-search to train and test different classifier algorithms with different parameters
    a. XGBoost
    b. Random Forrest
    c. Neural Nets
    d. AdaBoost
16. Create Binary Classifier for each product (i.e. with Random Forrest)
    a. Predict the probability for each product (0,1)
    b. Loop Across different products
    c. Extract probabilities from the classifier that the row is for each product
    d. Combine probabilities
    e. For each row, find the 7 largest probabilities—rank-order
    f. Compute MAP@7 for each customer in test set (subset from train)
    g. Tune parameters and try different models
17. If ML algorithms not working well, explore using aggregating based on a feature

Works Cited:

Data Quest. “How to get into the top 15 of a Kaggle completion using Python.” dataquest.io. Web 2 May 2016. Available at: https://www.dataquest.io/blog/kaggle-tutorial/