Predictive Analytics Capstone

Combining Predictive Techniques

Introduction

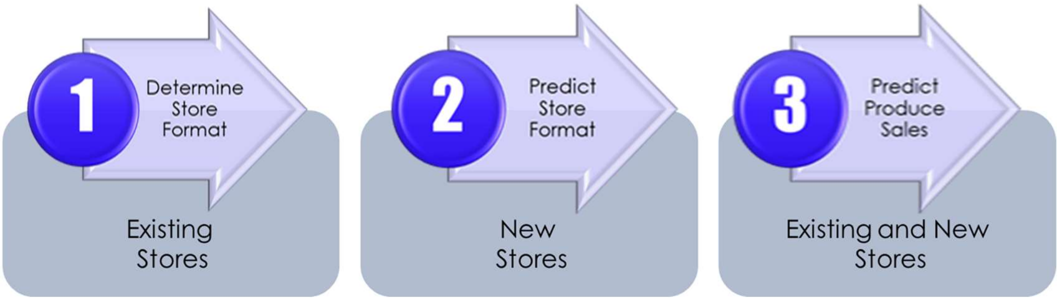
Project Rubric, https://review.udacity.com/#!/rubrics/437/view
Nanodegree, https://www.udacity.com/course/predictive-analytics-for-business-nanodegree--nd008

The Business Problem

The company where I currently work has 85 grocery stores and is planning to open 10 new stores at the beginning of the year. Currently, all stores use the same store format for selling their products. Up until now, the company has treated all stores similarly, shipping the same amount of product to each store. This is beginning to cause problems as stores are suffering from product surpluses in some product categories and shortages in others. I have been asked to provide analytical support to make decisions about store formats and inventory planning.

Details

I have been asked to complete three tasks, which are the following and detailed in sequence.



Three data files were provided.

- *StoreInformation.csv*
  This file contains location data for each of the stores.

- *StoreSalesData.csv*
  This file contains sales by product category for all existing stores for 2012, 2013, and 2014.

- *StoresDemographicData.csv*
  This file contains demographic data for the areas surrounding each of the existing stores and locations for new stores.

Task 1: Store Format for Existing Stores

To remedy the product surplus and shortages, the company wants to introduce different store formats. Each store format will have a different product selection in order to better match local demand. The actual building sizes will not change, just the product selection and internal layouts. The terms "formats" and "segments" will be used interchangeably throughout this project.

I have been asked to:

1. Determine the optimal number of store formats based on sales data:
   - Sum sales data by StoreID and Year
   - Use percentage sales per category per store for clustering (category sales as a percentage of total store sales).
   - Use only 2015 sales data.
   - Use a K-means clustering model.
2. Segment the 85 current stores into the different store formats.
3. Use the StoreSalesData*.csv* and *StoreInformation.csv* files.

Task 2: Store Format for New Stores

The grocery store chain has 10 new stores opening up at the beginning of the year. The company wants to determine which store format each of the new stores should have. However, we don’t have sales data for these new stores yet, so we’ll have to determine the format using each of the new store’s demographic data. Determine the optimal number of store formats based on sales data.
I have been asked to:

1. Develop a model that predicts which segment store falls into based on the demographic and socioeconomic characteristics of the population that resides in the area around each new store.
2. Use a 20% validation sample with *Random Seed* = 3 when creating samples with which to compare the accuracy of the models. Make sure to compare a decision tree, forest, and boosted model.
3. Use the model to predict the best store format for each of the 10 new stores.
4. Use the StoreDemographicData.csv file, which contains the information for the area around each store.
5. Note: In a real world scenario, you could use PCA to reduce the number of predictor variables. However, there is no need to do so in this project. You can leave all predictor variables in the model.

Task 3: Forecasting Produce Sales

Fresh produce has a short life span, and due to increasing costs, the company wants to have an accurate monthly sales forecast. I have asked to prepare a monthly forecast for produce sales for the full year of 2016 for both existing and new stores. To do so, follow the steps below.

Note: Use a 6 month holdout sample for the TS Compare tool (this is because we do not have that much data so using a 12 month holdout would remove too much of the data).

1. To forecast produce sales for existing stores you should aggregate produce sales across all stores by month and create a forecast.
2. To forecast produce sales for new stores:
    - Forecast produce sales (not total sales) for the average store (rather than the aggregate) for each segment.
    - Multiply the average store produce sales forecast by the number of new stores in that segment.
    - For example, if the forecasted average store produce sales for segment 1 for March is 10,000, and there are 4 new stores in segment 1, the forecast for the new stores in segment 1 would be 40,000.
    - Sum the new stores produce sales forecasts for each of the segments to get the forecast for all new stores.
3. Sum the forecasts of the existing and new stores together for the total produce sales forecast.