Michael Surdek

Final Project

Speaker Notes

Bird Books LLC is a large online retailer that is launching a next generation e-Reader called Blue Jay in the first quarter of 2021. The company has seen significant growth over the past few years and wishes to capitalize on that growth by maximizing their marketing efforts for this new product. Using data on our customers, it is possible to predict when each one will be most likely to purchase the new e-Reader, based on sales of the previous generations of e-Readers.

There are two main data sources with which the Bird Books LLC is currently working. The first data set contains customers who have bought previous generations of the e-Reader. The data set includes some basic information as well as the customer’s browsing and purchasing history. The second data set contains current and potential customers. The goal is to place these customers into four categories of adopters based on trends that can be identified within the training data. The variables are similar across the two data sets. The variables gender, age, marital status, web activity, browsing and purchasing history, and payment method can be found in both sets of data. The training data set also includes the variable eReader\_adoption, which shows the category of adopter for each customer at the time of their previous purchase.

The four adoption categories are innovator (first week), early adopter (second or third week), early majority (fourth through eighth week), and late majority (after two months). This variable is important because it makes it possible to identify which variables affected purchase timing in the past and to predict when future customers will make their purchase. Although the data set contains many variables that will likely be solid predictors of purchase timing, some other data could be collected by the company that might also be useful. Examples of this type of data are home address, occupation, family size, and additional services that the customer uses.

With all of this data, it should be possible to exploit analytics in order to add business value and uncover new opportunities. The main goal of Richard’s project is to assign a “probable adoption time” to each potential customer. When this is done, marketing can be timed and targeted for each individual when they are most likely to respond. For example, customers identified as innovators might not require any marketing efforts. It is possible that these customers are going to purchase the new e-Reader within the first week no matter what. Instead of using marketing resources in this way, the company might decide to target the early adopters and turn them into innovators by offering a pre-order bundle. These are already good customers, who would likely buy the e-Reader within a couple of weeks, but it could be worth it to get them committed as soon as possible. Additionally, customers in the early or late majority might need more incentives to get over the hump. Marketing could target these customers a few weeks after the e-Reader is released and offer them benefits such as a free trial or additional services. Finally, the value of this project can be extended beyond the marketing of this one new e-Reader. The company can use insights derived from this project to increase web activity or to connect customers to other products at the optimal times with a “customers also bought” or recommended product applications.

The main goal of this initiative is to classify potential customers into 4 adoption categories to predict when they will purchase our new e-Reader. Each potential customer should be able to be considered an innovator (first week), early adopter (second or third week), early majority (fourth through eighth week), and late majority (after two months). With this information, we should hopefully be able to maximize our marketing efficiency by targeting marketing to each group of customers at the time(s) when they are most likely to respond. This initiative should have various organizational benefits. This project will attract new customers, encourage upgrades for existing customers, and connect them to our other services, leading to business growth in the short and long term. This initiative will be determined to be a success based on a couple of measurements. First, we can compare e-Reader sales of new generation vs previous generations. If more customers purchase the new e-Reader, it might be possible to say that the initiative was a success. Although the business has grown since the last e-Reader was released, we can control for that growth and identify whether the targeted marketing had an impact on overall sales. Another way to measure success of this initiative is to analyze individual customer’s predicted buying time vs actual buying time to see if targeted marketing was effective. The predicted buying time should have been based on past data which means it was the predicted time before applying any marketing effort. If many customers end up purchasing the e-Reader earlier than predicted, it might be safe to say that the marketing efforts paid off.

The four adoption categories are innovator (first week), early adopter (second or third week), early majority (fourth through eighth week), and late majority (after two months).

The Data Analytics Lifecycle is a key guide and process that we will follow throughout the project. In the discovery phase, we will analyze the project needs and frame the problem in its business context. In this case the Null Hypothesis is that all customers can be expected to purchase the new e-Reader at the same time. On the other hand, the Alternative Hypothesis is that some customers can be expected to purchase the new e-Reader at different times than others. In the data prep phase, we will turn the data that we have into the data that we need. This includes asking questions such as what data might predict buying time? Do we have this information? How can we get it? In the model planning phase, we will try to determine what method is best for this analysis. This could be a t-test, linear regression, ANOVA, or something else. In this case, the best method might be a chi-squared test, if most of the data ends up being categorical in nature. The next phase, model building, is when we will use the training data to build model and the testing data to evaluate it. Since we have all of the variables that we think might be useful, this phase is the process of working through each one to build the best model to predict customer buying time. The next step is to communicate the results. The model’s results and finding will need to be presented to all key decision makers as well as the marketing department who will be deploying strategies based on the model. This leads directly to the operationalize phase, where we will collaborate with marketing to create a plan which will maximize marketing efficiency of the e-Reader. The marketing efforts could include targeting the innovator group very early by offering a pre-order bundle. Additionally, the early majority might need more incentives such as a free trial or additional service offerings.

Following the data analytics life cycle will lead to the best possible results of this project when all is said and done. Each phase has benefits that will help our results be more predictable, reliable, and secure at the end of the day. The discovery phase ensures that there is a problem that needs to and can be solved and that we will use the optimal resources. The data prep phase ensures that the data we have is the best data to represent our population of potential customers, increasing reliability. The model planning phase ensures the statistical analysis can provide information that might be able to solve the problem. The model building phase optimizes performance and quality of the statistical analysis. When we communicate our results, it improves clarity of message, aligns objectives across departments, secures information and business practices. In the operationalize phase, we put our ideas into practice and watch the benefits come to life.

The data that is available at this time is a database of our current and potential customers that includes basic info, browsing & purchasing history. This is quality data because it covers a wide variety of customers over time, large sample. On top of that, the trends that we identify will hopefully be relatively apparent and stable over time. This means that there should be particular variables that clearly affect each customer’s buying time. There are also limitations to the data as it currently stands. The largest limitation is that there are likely many currently unavailable variables that could be more predictive than what we have. Potential variables I have considered that might be more predictive are home address, occupation, family size, and additional services that they use on our website. This leads to the questions of how can we get that information from our customers, and if it is available through the internet, can we obtain it legally?

We have attempted to use data analytic tools in the past to predict the purchases that customers will make. We have used association to discover relationship between items that individual customers might purchase. With this tool we have found out that if a customer purchases certain items, they are more likely to buy other specific items. We use this information to display items that are frequently bought together. We have also used a random forest approach to try to determine if there are things that we know about our customers that can predict what they will buy and when. Through past analytic initiatives, we have discovered that information such as income and family size provide insights into what customers might be looking for.

With past generations of the e-Reader, we have used mass marketing campaigns to advertise the new products leading up to and beyond the release dates. All customers who had not yet purchased the e-Reader or unsubscribed from updates would receive emails on pre-determined dates. The response rate and conversion rate for these campaigns have been below industry averages, which is why we would like to switch to a targeted marketing strategy.

These previous analytical methods are helpful but they are not exactly what we need for this initiative based on the data that we currently possess. Association is an inefficient tool because it requires a lot of time input which grows exponentially as the data becomes more complex. Association will also not work for this initiative because not all of our data is categorical, which is what you need to determine the associations.

Our previous marketing campaigns have also shown glaring weaknesses, on top of the below average response and conversion rates I have mentioned. Sending emails at pre-determined times causes us to miss customers who might have been ready to buy for only a brief period of time. Additionally, sending these emails in the weeks leading up and immediately after release to certain customers who are more likely to be in the early or late majority might cause them to become fatigued by our message and discourage them to make the purchase when they would normally be ready.

For this initiative, the 2 possible methods we can use are classification and clustering. Classification is when you assign labels to objects. This method makes sense in this case because the objective is to categorize our customers into the 4 adoption categories of Innovators, Early Adopters, Early Majority, and Late Majority. The other method, clustering, is when you group items by similarity. This fits our situation because customers that are more similar to each other might be more likely to fit into any one specific adoption category. This method also has additional benefits because it is easy to implement and easy to build on. The clustering method would provide a foundation of analysis that we could scale towards future product offerings.

These are 4 examples of customers that might fit into each of the 4 adoption categories. The information listed is basic information about each of them which was used along with other information to make the classifications.

We were able to use past data to predict when each of the 4 customers would be most likely to purchase our e-Reader if we rolled it out like normal and did not apply any targeted marketing efforts. Using research on marketing campaigns, we were also able to predict when each of the 4 customers would be most likely to purchase the e-Reader if we roll it out and target marketing towards each individual when they would be most likely to respond positively. For example, David is an early adopter who would normally wait a week before purchasing the e-Reader. He likes to make sure customer reviews show the product is as expected and delivers on its promises. With targeted marketing such as a sale over the first few days on the marketing, David is more likely to purchase the e-Reader after just a few days. He would not pre-order the e-Reader like an innovator would, but he would still be willing to commit for a discount once he sees actual customer reviews.

The main value of this initiative which makes customers purchase earlier than usual is that it will increase overall sales by lessening the negative impact of time on the likelihood a customer will purchase. When a product is introduced, you can map the likelihood that any individual customer will eventually make the purchase. Since all future days are incorporated into the overall likelihood, as time goes on and days pass, the likelihood only decreases. By convincing customers to commit to purchases earlier, more e-Readers will be sold in total.

This data analytics initiative will have a multitude of benefits for our organization. The first, as mentioned above, is increasing total sales. The best question to ask yourself is how many potential customers do we typically lose with each passing day? Many of these customers are lost simply because time passes, and our product is no longer on their mind like it was before. When you look at it this way, this initiative has clear value. Additionally, customers making earlier purchases has benefits beyond just increasing overall sales. Customers who purchase earlier will have a larger customer lifetime value (CLV). This is because they are more likely to purchase our future e-Readers, and also because we will gain more revenue from the more time they use our other products and services. Another benefit is referrals. Customers who make earlier purchases will get the most out of their products and be more satisfied with the product and our organization. This leads to word of mouth referrals, which is possibly the most effective marketing strategy that exists. Finally, this data analytics initiative will create insights and show trends that can be applied to future product releases and more extensive projects.