DATA SCIENCE PROJECT REPORT

Prepared By:
Data Detectives
Benjamin Bloch (bb1976),
Titash Mandal (tm2761),
Sanket Nawle(skn288),
Jayesh Patil(jpp421)

INTRODUCTION

The United States pharmaceutical industry is a cut-throat world full of companies each battling for prescription drug supremacy. Vast amounts of resources are poured into the research and development of drugs that aim to alleviate many of the most challenging and complex diseases Americans face. One major key for the biotech bottom line is capitalizing on a drug that finally makes its way through the trial and FDA approval process and lands on the shelves of pharmacies for retail sale. These brand name drugs, with patents that protect them from competition for decades, are the profit drivers for pharmaceutical companies. Eventually, however, the patent on the drug expires, and generic drugs that claim to be just as effective at treating the respective disease enter the market, giving consumers and doctors lower cost options that can erode the profits of brand-name drug makers. **In the last year, the public outcry against high drug prices has reached a frenzy, with Congress and the President both pledging to work on legislation to help bring down drug costs for consumers.** In the face of rising competition for their brand name product, as well as increased clamoring by consumers to lower costs, pharmaceutical companies must find ways to continue effectively selling their brand name drugs.

Generic drugs must prove in trials to the FDA that they are similarly effective to the brand name alternative - they must use the same active ingredient as the brand name drug, and they must act "**at the same rate, and to the same extent**" as their brand name counterparts (thedenverchannel.com). Despite these regulatory guidelines, not all doctors are convinced that generic drugs are as effective at treating illness as brand name drugs. In a recent article published on the marketplace.org website, news reporter Tracy Weber reported her findings that despite the existence of generic options to treat a particular disease, doctors were still prescribing brand name drugs because they feel that the brand name option is more effective (marketplace.org). An article published in the Washington Post quotes a survey published in the Annals of Pharmacology that reports nearly 50 percent of doctors hold some negative perception about the quality of generic medications (washingtonpost.com). There may be other, more nefarious reasons why doctors are prescribing brand name drugs however that have nothing to do with patient interests. The New York Times published an article in August of 2017 that details ways in which the pharmaceutical industry incentivizes doctors to prescribe brand name drugs so that they can boost profits. Doctors may be motivated to prescribe brand name drugs instead of generics by the possibility of increased profits in the case of uninsured patients (nytimes.com). Even in the case of insured patients, doctors are wined and dined by pharmaceutical companies and may be influenced to use their drugs because of those efforts. A report based on government data published in 2014 shows that more than 500,000 physicians in the U.S have ties to the pharmaceutical industry (consumerreport.org). An extensive study conducted by ProPublica demonstrated that doctors who received money from drug and device makers prescribed a higher percentage of brand name drugs overall than doctors who didn't get the payments (npr.org).

MOTIVATION

There been a lot of debate already about the use of generic drugs vs. the brand names. **We expect a doctor's prescribing behavior to be governed by many complex, interacting factors related to the person's specialty, the city he is located in, his personal preference, the research he does, and so on.** According to the Huffington Post, the use of generic drugs is expected to grow over the next few years as a number of popular drugs came off patent in 2015. Despite the prevalence of generic drugs, there are still many who are prescribing brand name drugs. California doctors were among the top prescribers of brand name drugs, the Los Angeles Daily News points out. "Almost 65 million Medicare Part D claims totaling $7.5 billion were filed by internal medicine, family medicine & general practice physicians from the Golden State for their patients," the newspaper reported (mercurynews.com). Such headlines beg the question: Why would doctors prescribe more expensive drugs when clearly cheaper alternatives are available? **In an effort to learn more about the healthcare system we decided to explore the medical dataset to understand the behavior of the prescriptions and find reasons for such behavior.**

SUPERVISED LEARNING PROBLEM

Our research team set out to build a supervised predictive model that could pinpoint doctors who are ripe for pharmaceutical companies to market their brand name drugs effectively. Whether they are motivated to prescribe brand drugs due to their belief in the drug's superiority or due to deals with pharmaceutical companies, our goal is to use specific demographic and medical provision data to predict the likelihood of a doctor prescribing a high ratio of brand name drugs to generic drugs. Our model will help focus pharmaceutical companies' marketing efforts. There are over one million doctors in the United States (statista.com) and a targeted marketing approach can ensure that pharmaceutical companies market their drugs to the doctors most likely to prescribe them. **Researchers estimate that 73 billion dollars per year are earned by brand name drug prescribers in cases where there is a generic equivalent (clark.com)**. Ensuring that pharmaceutical companies receive a large share of that will help pad company profits.

The data we used to train and test our model comes from the **U.S Centers for Medicare and Medicaid Services (CMS).** This dataset contains all medical claims filed under the Medicare Part D prescription drug plan for each calendar year. Our team used the dataset from 2013, which was also used by Roam Analytics to build a predictive model.

However, the Roam study focused on a different question than our team studied, namely: Can one predict a doctor's demographic features based on their specific prescription history? They used a sparse matrix of drug names as features, and rotated through a series of demographic target variables (roamanalytics.com). **In our case, we wanted to study the opposite question: Can a doctor's demographic features predict if she will prescribe a brand name drug?**

Understanding the Data

The CMS dataset has a rich array of data, containing 84 features and just over 1 million rows. Each row, or data instance, represents one doctor who prescribed medication to at least 11 patients in 2013. The large size of the dataset will allow us to be relatively confident that we will not have a large variance error in our model, since sample size is a key component to minimizing error due to variance. The large sample size can thus help control overfitting, in addition to the use of cross validation techniques.
Due to the nature of the dataset, selection bias is inherent in our models. Most participants of Plan D Medicare drug program must be 65 and older (although there are some exceptions), and so our sample is not representative of all medical patients in the U.S. By recognizing the existence of this sampling bias, we can still obtain useful predictive models, so long as we focus our predictions on the population of patients 65 and older. Most elderly Americans use Medicare insurance, so this dataset is a good sample of the 65 and older population.

The dataset contains two key fields that, when combined, became our target variable. **These fields are the number of brand name prescriptions filled for a particular doctor, and the number of generic prescriptions filled for that doctor.** Our target ratio, termed "**brand ratio**", became the number of brand prescription counts divided by the total number of prescription counts for a particular doctor. By using a ratio, we ensured that our target variable was on a scale that could generalize across doctors, some of whom may have had a smaller caseload than others. However, generating our target variable wasn't as straightforward as it may seem, as will become clear in the following paragraphs. The mean of the brand ratio was 0.23; a doctor prescribes a brand-name medication 23% of the time.

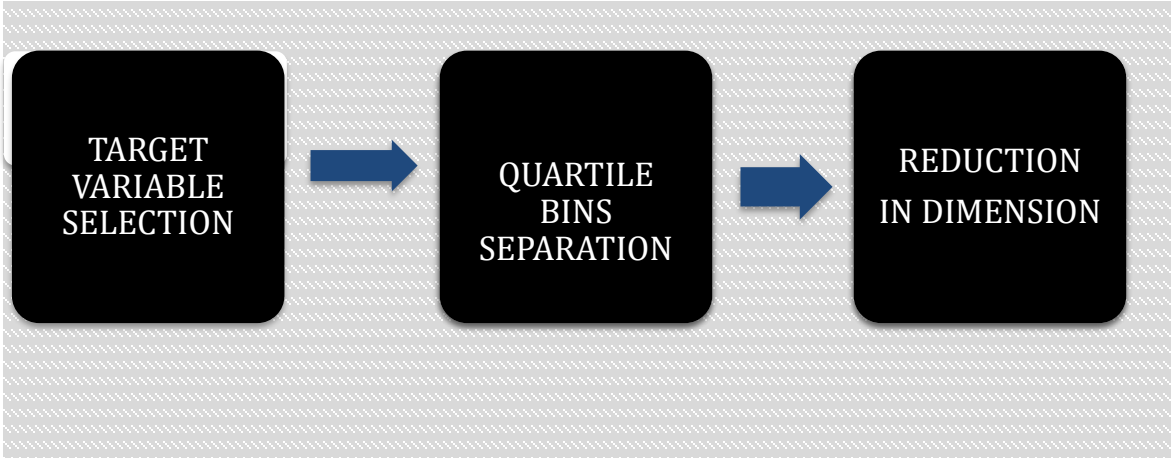


Fig-1: Figure showing steps taken towards data cleaning.

The **brand ratio is a continuous variable**, which makes models such as linear regression the ideal candidate for prediction. However, we also wanted to test classification models to see how they performed, and so we created quartile bins of the brand ratio for use in multi-label classification. Each of the four bins represents 25% of the data, and the split was made based on the four quartiles of the brand ratio. These quartile bins are good splits because they are a large enough number to give some variability to the classification labels, but are not too many splits that would result in low model accuracy.

DATA CLEANING AND FEATURE ENGINEERING

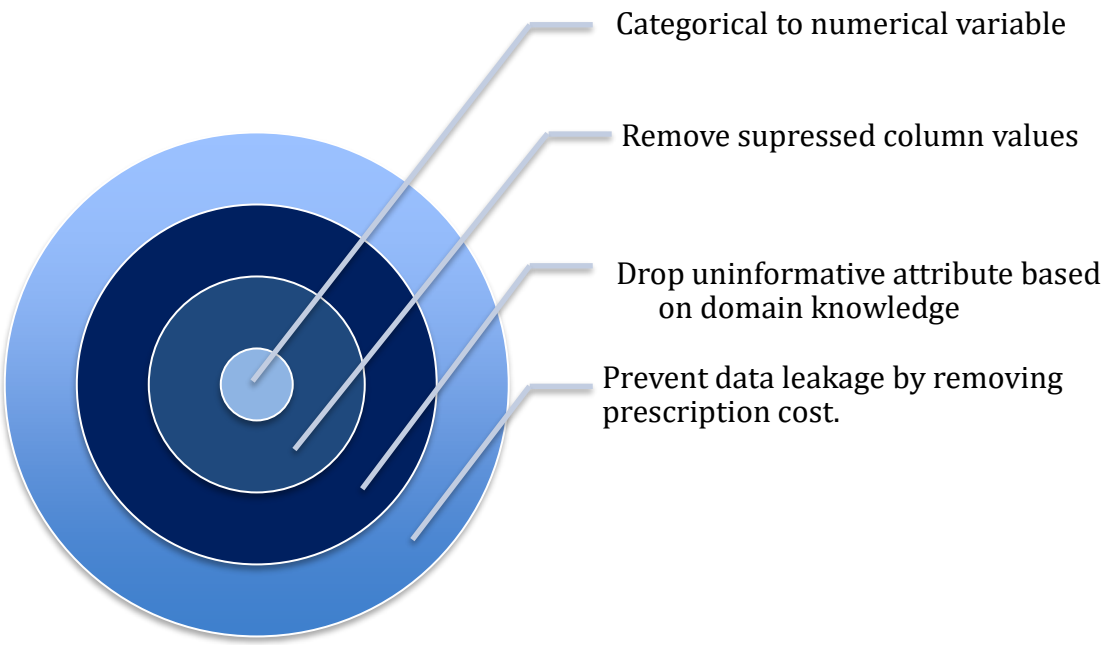


Fig-2: Figure showing the steps towards feature engineering

Much of our time and resources were devoted to data preparation and cleaning. We well understand the axiom of "**garbage in, garbage out**", and we allocated about 80% of our time to working on improving our dataset before running it through our models. Some of the data cleaning and preparation challenges we encounter are as follows:

Due to the sensitivity of medical data and the regulatory need to protect patient privacy, **CMS suppressed a number of key fields in their release to the public.** As a result, despite the large number of rows in the data set, **only 51% of the instances** had revealed values for brand claim counts and generic claim counts, the two key fields we needed to generate our target variable. We needed to devise a methodology for imputing missing values, since we did not want to lose half our data instances, especially because the data was not missing at random. Fortunately, CMS included an indicator variable that gave us guidance on whether the suppressed value was greater or less than 10. We used that guidance to fill in missing values where necessary using a consistent placeholder value for the instances with under 10 counts, and then used that placeholder to calculate the missing values for the other fields (a more detailed explanation is contained in the appendix). In order to check what the effect of estimating these missing values was, we ran our models on two datasets, one that dropped all rows with missing data, and one that used our estimation methodology to retain the rows with missing data. The mean of the brand ratio target variable for the dataset with estimations was 0.19, compared to 0.23 for the dataset without estimations.

Additionally, there were many other features intrinsic to our analysis where there was missing data. These cases were not due to suppression by CMS, and in those cases, we imputed the missing values using averages calculated on the data that we had for the respective feature.

The dataset contained a mix of categorical and numeric variables. There were 6 categorical variables in the dataset representing key demographic information. Since these categorical variables were unordered, or nominal, we transformed them into a matrix using One **Hot Encoder** for use in modeling. Extensive

cleaning was performed on string variables so that proper aggregation by One Hot Encoding could be accomplished.
A number of feature engineering procedures were performed on numeric features as well. **Similar to how we generated the target variable, a number of features were transformed to ratios so that they could be generalizable regardless of the number of patients a doctor saw.** Examples of features that we transformed in this manner are **patient gender and patient race**. The feature "**beneficiary count**" was not transformed so we could retain a feature that could capture the impact of high patient counts versus low patient counts on prescription behavior.

Additionally, we dropped many of the 84 features in the original dataset before modeling, due to our domain knowledge that they would not have an impact on prescription behavior.

We were cognizant about the concept of leakage, and did not leave any features that would act as a stand-in for our target variable for easy prediction. **As such, we dropped variables such as total cost of prescriptions, since brand name drugs have a higher cost and therefore a high total cost would allow the model to easily predict a high brand ratio.**

We **split our training set into two parts (75% to training, and 25% to test) in an effort to establish a test set that we could use to test our models.** We did not touch this test set until our training model was completely tuned so as to avoid leakage from the test set into the training set.

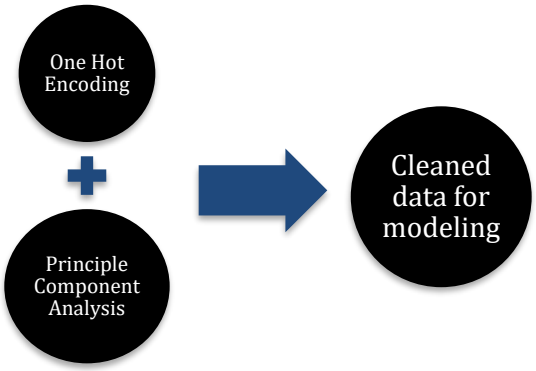


Fig-3: Figure showing conversion from categorical to numerical variables removing bias.

One Hot Encoding helped convert categorical features into a binary matrix that we could use for modeling, but it also created a final dataset with over 300 features. This caused our models to run slowly, and in some cases to crash entirely due to the size of the feature space and the size of the data set. **We therefore chose to implement two feature reduction methodologies: Principle Component Analysis (PCA) and Tree-Based Pruning**.

For PCA, **we tested different thresholds of components** to locate where increasing the feature space did not provide us with a significant increase in accuracy, and we chose that number of components to leave in our model. **29 components accounted for 93% of the variance in our dataset of 343 individual features.**
The downside to using PCA is the linear transformation left us with features that were not interpretable anymore; we could no longer identify what individual features were the key drivers of information gain. To mitigate this, we also reduced dimensionality on the original dataset by pruning features that did not contribute much information gain using a tree based classifier and the **SelectFromModel in Scikit Learn**. This helped reduce our feature space while also retaining the original features so we could identify which features were predictive.
Before implementing any models, we tuned the hyper-parameters using cross-validation techniques. Each model was run using 3-fold cross-validation, and optimal hyper-parameters were chosen based on the maximum mean of the accuracy on the three folds. By using cross-validation, we obtained the best combination

of hyper-parameters that would maximize accuracy while controlling for overfitting. This was particularly key when we implemented decision trees, which if left unchecked, will over fit to produce a 100% accuracy on the training set.

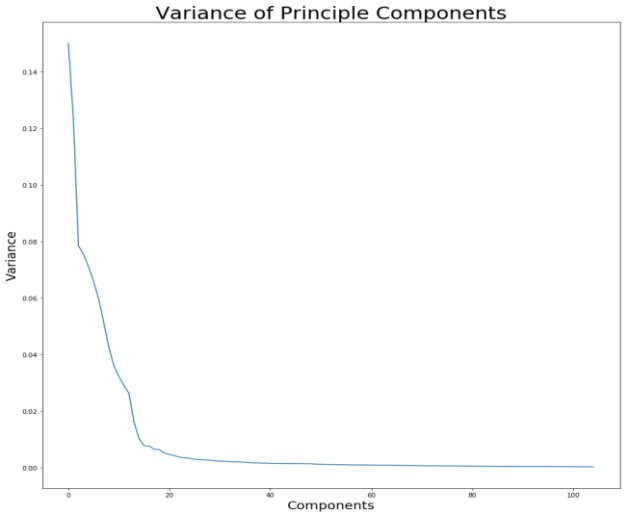
Before implementing and running any models, we set an evaluation baseline so that we could establish a benchmark for our models to improve upon. For any classification models, we simply wanted to beat a model that randomly guesses one of the four quartile classes every time, which would achieve an accuracy of 25%. For regression models that used a continuous target variable, we converted the model's brand ratio predictions back into quartile bins, and then compared those predicted quartiles to the true quartile bin values to obtain an accuracy that we could compare to our classification models. Due to the complex nature of multi-label classification and the need to compare across both classification and regression models, we decided that straight accuracy would be a good starting place for evaluation. Additionally, since there are so many doctors in the United States and companies can only choose to market their products to a small subset of them, the need for an accurate model is crucial.

Quartile Ranges for Target Variable (Brand Ratio)

Dataset	**0-25%**	**25-50%**	**50-75%**	**75-100%**
Data	0.0 to 0.135	0.136 to 0.204	0.205 to 0.282	0.283 to 1.0

Fig-4: Graph showing the change of variance of attributes in the dataset

The first item to point out is that our models yielded better accuracy scores across the board when implemented on the dataset that did not contain any estimations for missing values. Our initial reaction to this outcome is that the dataset that did have estimations in it had a higher percentage of rows where there were low prescription counts. These low count prescribers were hard for the model to estimate, since they did not have a lot of prescription data that could serve as a significant sample size. In effect, removing these doctors from the dataset helped control for outliers who would otherwise confuse the model due to their small prescription size. The other possibility is that our estimations for missing data were incorrect or biased in some way, and that caused the models to become less accurate. Further study is needed to assess what the true cause of the accuracy decrease was, and we will be exploring this further in a follow up project. Since our model accuracy was better on the dataset without any estimations, all discussions regarding evaluation will focus on the results from that dataset.



DATA MODELING AND ANALYSIS

Our only non-classification model was **linear regression**. Linear regression can be implemented on continuous variables, and since our brand ratio was continuous before we binned, we wanted to see if Linear Regression would be a good model to use for prediction. We tested three types of linear regression models to see which one handled predictions the best: **Ordinary Least Squares, Ridge Regression, and Lasso Regression**. The results for all three models were relatively similar, with r2 scores hovering right around 0.54. OLS was the best performer, with an r2 of 0.547.

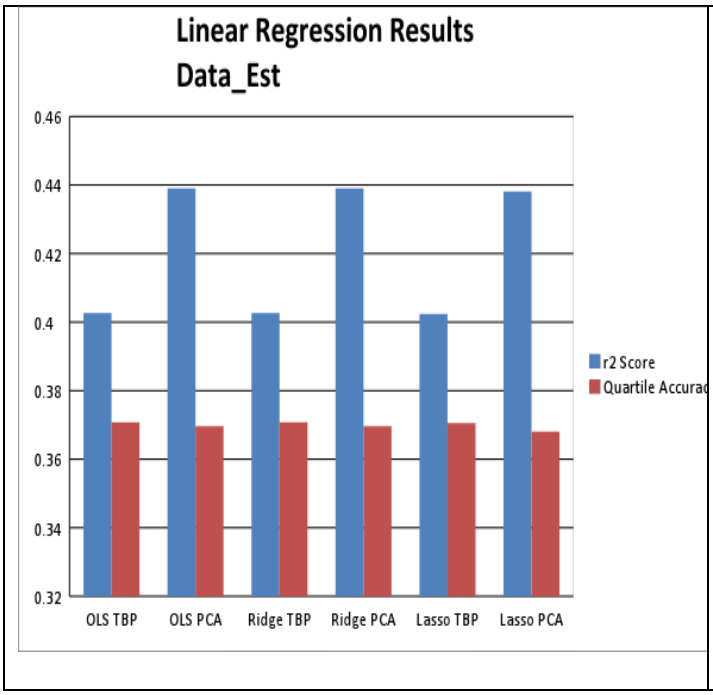


Fig-5: Linear regression results from data_est

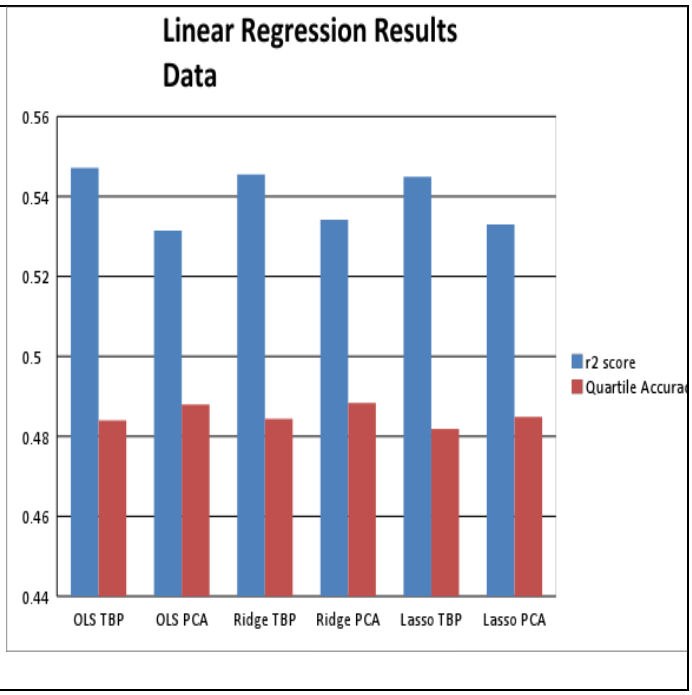


Fig-6: Linear regression results from data

In general, the models that used tree based pruning did about 0.01 better on the r2 score than the models that used PCA. All models also performed similarly when converted into quartiles and tested for classification accuracy, achieving accuracies around 48%. While it is a bit surprising that all three models performed similarly, we think that using PCA before implementing the models helped put them all on an even playing field. Ridge and Lasso are designed to help reduce multi-collinearity that OLS typically suffers from, and PCA also helps reduce the effects of multi-collinearity. In fact, some of the beta coefficients from the models that did not use PCA as a preprocessing step were outset the typical 1 to -1 range, indicating that there were some features that had high multi-collinearity.

What is interesting to note about the differences between the models is the variability in feature importance. Many features exhibited similar coefficients across the three models, but there were some that were vastly different. **One good example is the state of Connecticut. In OLS, doctors in Connecticut exhibited a high beta coefficient, but in Ridge and Lasso, the beta for these doctors was much lower. This seems to suggest that the Connecticut feature is suffering from some multi-collinearity, which Ridge and Lasso help reduce**. Another interesting feature is the state of California. In our regression models, it emerged as a strong determinant feature, but that was not the case with our classification models. This difference also likely has something to do with linear regression's sensitivity to multi-collinearity.

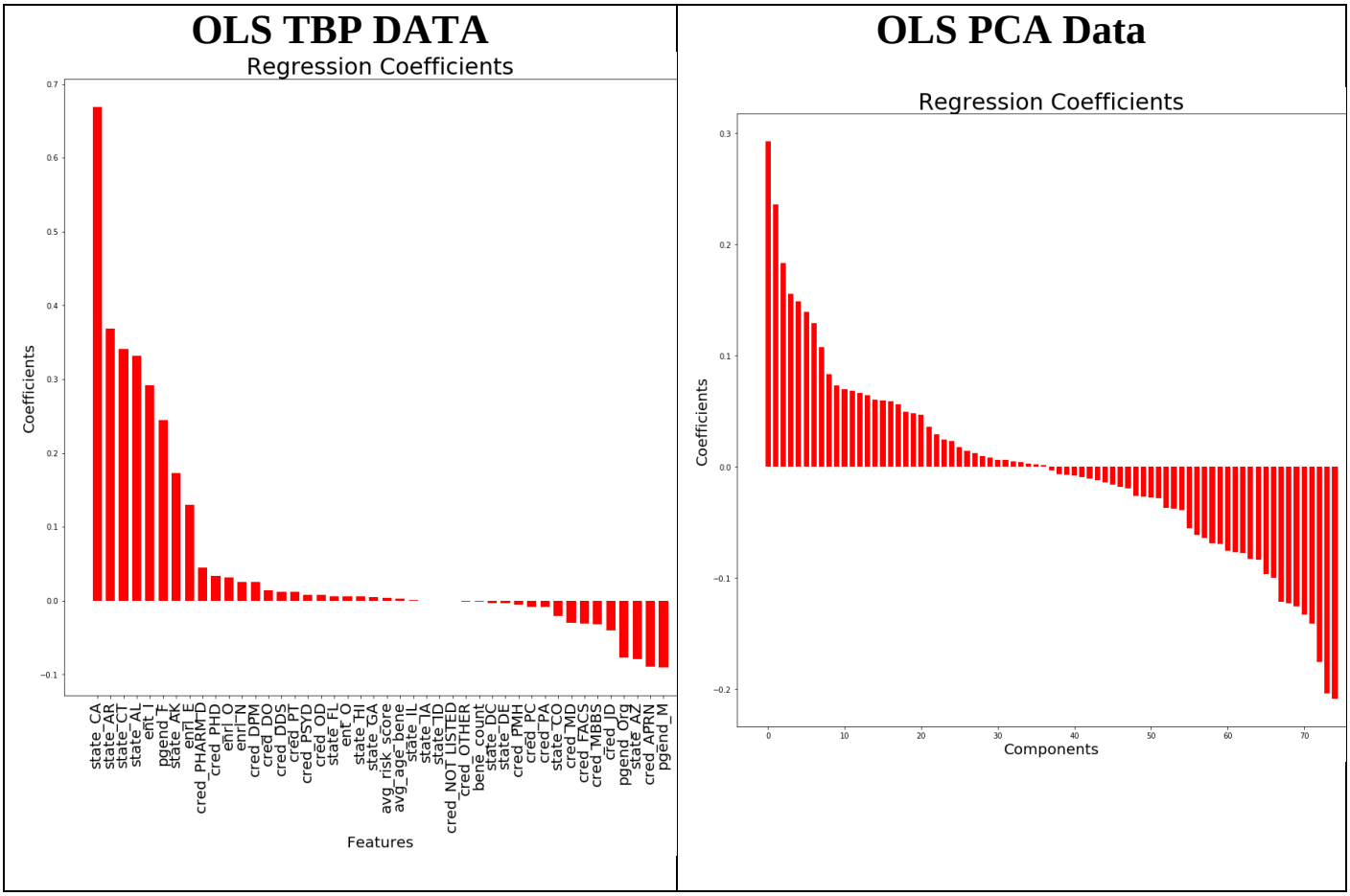


Fig-7: Graph showing the OLS TBP Data and OLS PCA Data regression coefficient

We then tried logistic regression. Unlike linear regression, logistic regression does not predict the actual value of a numeric variable given a training set. Instead, the output is a probability that the given input point belongs to a certain class. Thus, it acts as a classification model in our case to understand whether a doctor is inclined to prescribe generic or branded drugs given the required attributes. As seen from the results, we obtain similar training and testing accuracy for models with or without using PCA. This may have happened since logistic regression takes care of the weights of the attributes, i.e., give more weight to the important attributes while classification. Hence, using PCA doesn't improve the accuracy significantly. **Accuracy this model for using tree based pruning is 53 % and with PCA it is 51 %.**

Next we implemented a classification model: decision trees (DTs). DTs are a non-parametric supervised learning method used for classification. The goal of a decision tree is to create a model that predicts the value of a target variable by learning from simple decision rules and inferring from the data features. Our target variable was binned as quartiles, hence using a decision tree was appropriate. A major reason for using the decision analysis method is its ability to assign specific values to problems, decisions, and outcomes of each decision. This reduces the ambiguity in analyzing the large dataset available to us. Moreover, the reasoning process behind the decision tree model is clearly evident when browsing the tree. This is in contrast to other "black box" modeling techniques in which the internal logic can be difficult to work out.

However, in decision trees if we do not tune the hyper-parameters, over-fitting will occur where the tree is designed so perfectly that it fits all the samples in the training data set. Thus, it ends up with many branches and with strict rules of sparse data. This affects the accuracy when predicting samples that are not part of the training data set. When we tried to run our dataset on the training sample, the decision tree gave us a 99% accuracy due to overfitting. The result looked like something below.

After tuning the hyper parameters and implementing our fitted model on the test set, we achieved an accuracy of 53% for TBP and 50% for PCA. These results beat our linear regression model, and so we dove into an analysis of the important features highlighted by our decision tree model.

The plot of the results from the decision tree gives us a graph that ranks feature importance based on information gain:

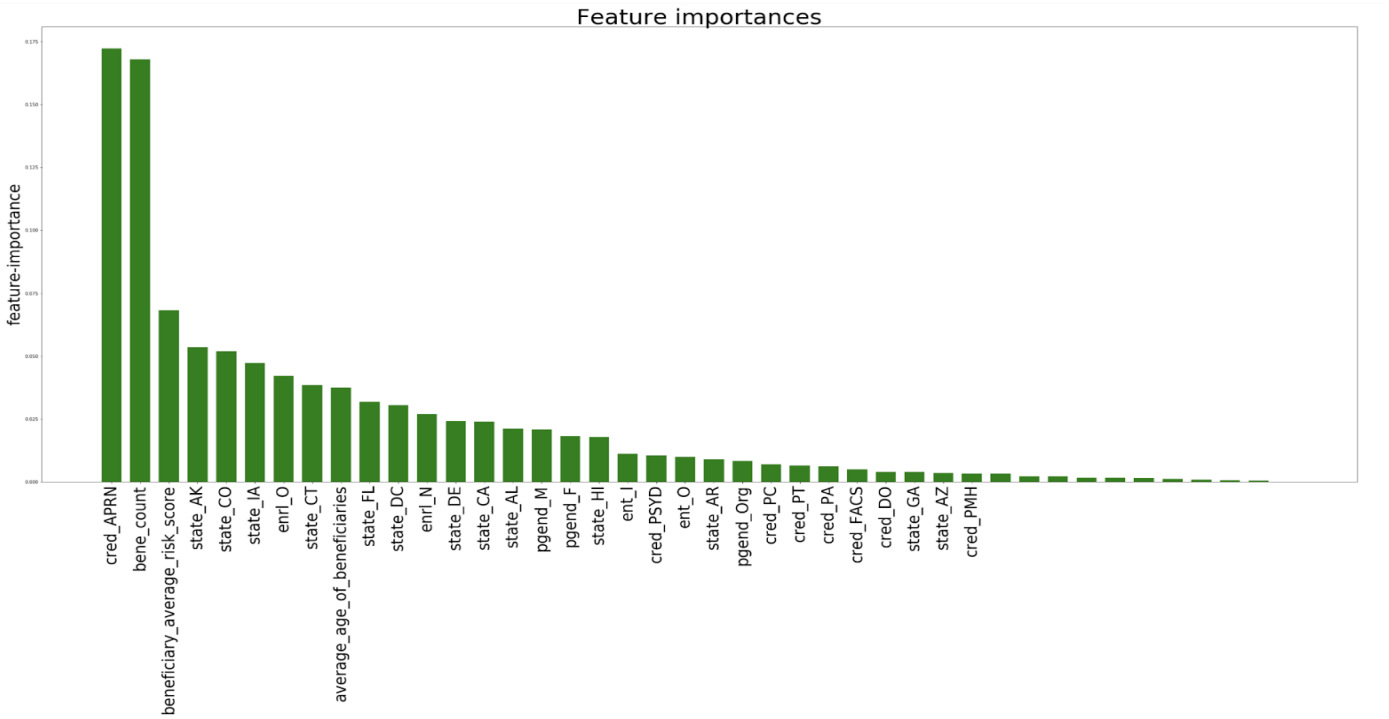


Fig-8: Graph showing the most important features after training a decision tree algorithm.

Thus, we can see that the most informative attributes are if a the medical provider was an advance nurse practitioner (cred_APRN), the number of unique patients a doctor saw in the year (bene_count), and beneficiary_average_risk_score, which is explained in more detail below. It is surprising to see that doctors practicing in Alaska (state_AK), Colorado (state_CO), and Iowa (state_IA) is also one of the most informative attributes. To know the relationship between these informative attributes and our target variable- brand_ration, we decided to take a look at the coefficient matrix. Correlations can shed light on what the direction of the relationship is, which decision trees cannot provide. Correlation will only be interpretable for the numeric features in our dataset however.

The attribute **cred_APRN** is a categorical variable, so we cannot use correlation as a measure for the direction of the relationship. But the linear regression model we ran indicates a highly negative beta coefficient for this feature, from which we can infer that advanced nurses practitioners are inclined towards generic drugs.

With **bene_count** being the second most informative attribute and having a positive correlation with our target variable, multiple conclusions can be drawn. One possibility could be that if a doctor sees many patients, he is extremely busy and may not have the time to think deeply about which generic medicine to prescribe, as there are many available in the market nowadays. Hence, he may be more inclined to prescribe a brand name medication because it is easier to remember. Money may also be a motivating factor for many doctors. He can prescribe the brand name of a particular drug if that company pays him a special commission to prescribe their medicine. More patients means more prescriptions which might mean a higher commission from the drug company. Additionally, having more patients may mean doctors desire a faster effective recovery of patients.

Due to an inherent trust on branded drugs some may want to prescribe them to their patients because they feel these drugs are more effective at providing a faster recovery time as a lot of money is invested by these companies when developing them. With better treatment, they feel they can maintain their high number of patients and not lose many patients altogether if the generic drug fails to achieve its promise.

The **beneficiary_average_risk_score** indicates the HCC risk score of a patient. HCC is a payment methodology based on "risk" used by CMS to adjust health plan payments at the patient level. The HCC risk adjustment is based on the enrollee health status and their demographic characteristics. As this has a positive correlation with our target variable, we can make a fair prediction that the doctors want to ensure that their high-risk patients do not lose time in their potential recovery, so they prescribe brand name drugs that they believe are guaranteed to work more effectively. Taking the health into consideration the doctors may feel that the overall result in patient care will be cost-effective and strongly benefit their high-risk patients.

When it comes to the geographical feature importance, we observed that doctors in Alaska and Colorado are major features. Doctors in Alaska have a strong positive relationship with brand name prescriptions, and doctors in Colorado have a strong negative relationship with brand name prescriptions. This may indicate particular state laws that encourage or discourage a doctor from prescribing brand name medication. Alternatively, perhaps the medical schools in these states have a particular philosophy regarding brand name drugs and that is what causes doctors in those states to have a particular orientation.

	cred_APRN	bene_count	state_AK	state_CO	state_IA	cred_PT	enrl_O	enrl_N	beneficiary_average_risk_score	brand_ratio
cred_APRN	1.000000	-0.048590	0.010033	-0.004075	0.008461	-0.006858	-0.016693	-0.028560	0.038847	-0.001329
bene_count	-0.048590	1.000000	-0.023416	-0.028980	0.008738	-0.014036	-0.037674	-0.313880	-0.020950	0.001923
state_AK	0.010033	-0.023416	1.000000	-0.005870	-0.004552	-0.000998	0.001325	0.000166	-0.005936	0.001363
state_CO	-0.004075	-0.028980	-0.005870	1.000000	-0.012845	-0.001470	0.002414	0.002077	-0.017216	-0.001992
state_IA	0.008461	0.008738	-0.004552	-0.012845	1.000000	-0.002183	-0.002032	-0.016035	-0.028777	0.004290
cred_PT	-0.006858	-0.014036	-0.000998	-0.001470	-0.002183	1.000000	-0.001447	0.011631	-0.000930	0.000748
enrl_O	-0.016693	-0.037674	0.001325	0.002414	-0.002032	-0.001447	1.000000	-0.033928	-0.039889	-0.000146
enrl_N	-0.028560	-0.313880	0.000166	0.002077	-0.016035	0.011631	-0.033928	1.000000	-0.053083	-0.002964
eneficiary_average_risk_score	0.038847	-0.020950	-0.005936	-0.017216	-0.028777	-0.000930	-0.039889	-0.053083	1.000000	0.000068
brand_ratio	-0.001329	0.001923	0.001363	-0.001992	0.004290	0.000748	-0.000146	-0.002964	0.000068	1.000000

Fig-9: Coefficent matrix of the most important features

RANDOM FOREST METHOD:

Random forest is simply a collection of decision trees whose results are aggregated into one final result. We used the random forest model because of its ability to limit overfitting without substantially increasing error due to bias. The way that Random Forests reduces variance is by training on different samples of the data. A second way is by using a random subset of features. In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run. Each tree is constructed using a different bootstrap sample from the original data.

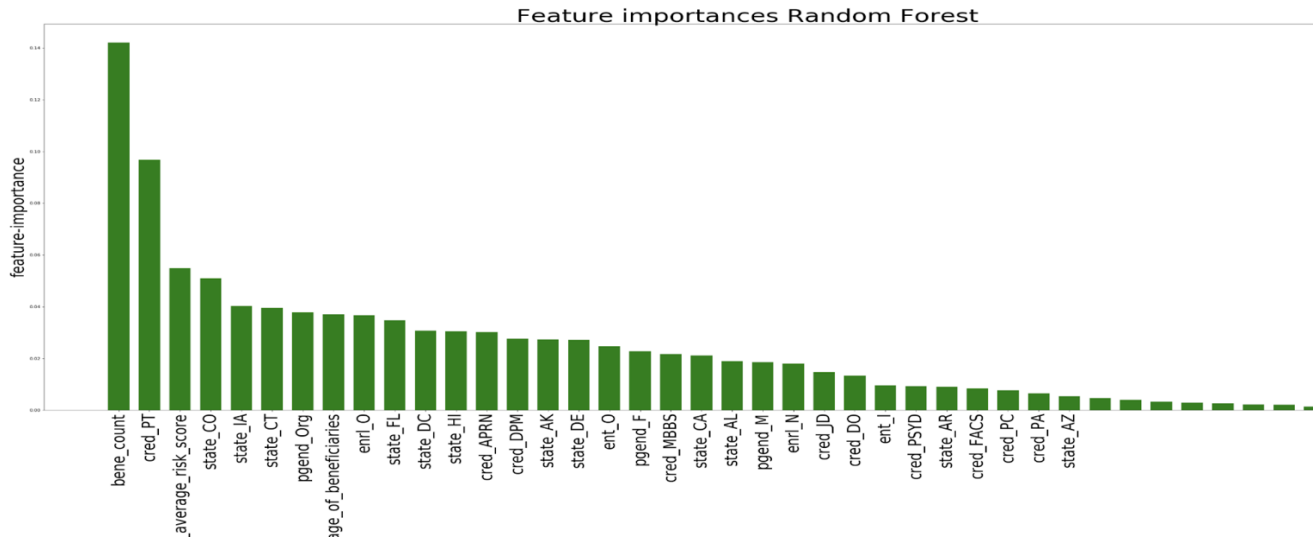


Fig-10: Graph showing the most important features after training a random forest algorithm.

```
Accuracy from the random forest model's test data set prediction (SELC) is 53.3224692681 %
Accuracy from the random forest model's test data set prediction (PCA) is 50.7515551984 %
```

Correlation between parameters in Random Forest

	cred_APRN	bene_count	cred_PT	state_CA	state_AZ	pgend_Org	state_GA	beneficiary_average_risk_score	brand_ratio
cred_APRN	1.000000	-0.048590	-0.006858	-0.044133	0.010497	-0.018607	-0.011863	0.038847	-0.001329
bene_count	-0.048590	1.000000	-0.014036	-0.007869	0.003865	-0.044129	0.019445	-0.020950	0.001923
cred_PT	-0.006858	-0.014036	1.000000	0.000457	0.000599	-0.001298	0.001609	-0.000930	0.000748
state_CA	-0.044133	-0.007869	0.000457	1.000000	-0.048727	0.026563	-0.053881	0.010350	-0.001383
state_AZ	0.010497	0.003865	0.000599	-0.048727	1.000000	-0.005198	-0.022051	-0.001558	-0.002364
pgend_Org	-0.018607	-0.044129	-0.001298	0.026563	-0.005198	1.000000	-0.003074	0.003616	-0.000464
state_GA	-0.011863	0.019445	0.001609	-0.053881	-0.022051	-0.003074	1.000000	0.005374	-0.002627
beneficiary_average_risk_score	0.038847	-0.020950	-0.000930	0.010350	-0.001558	0.003616	0.005374	1.000000	0.000068
brand_ratio	-0.001329	0.001923	0.000748	-0.001383	-0.002364	-0.000464	-0.002627	0.000068	1.000000

Fig-11: Coefficent matrix of the most important features

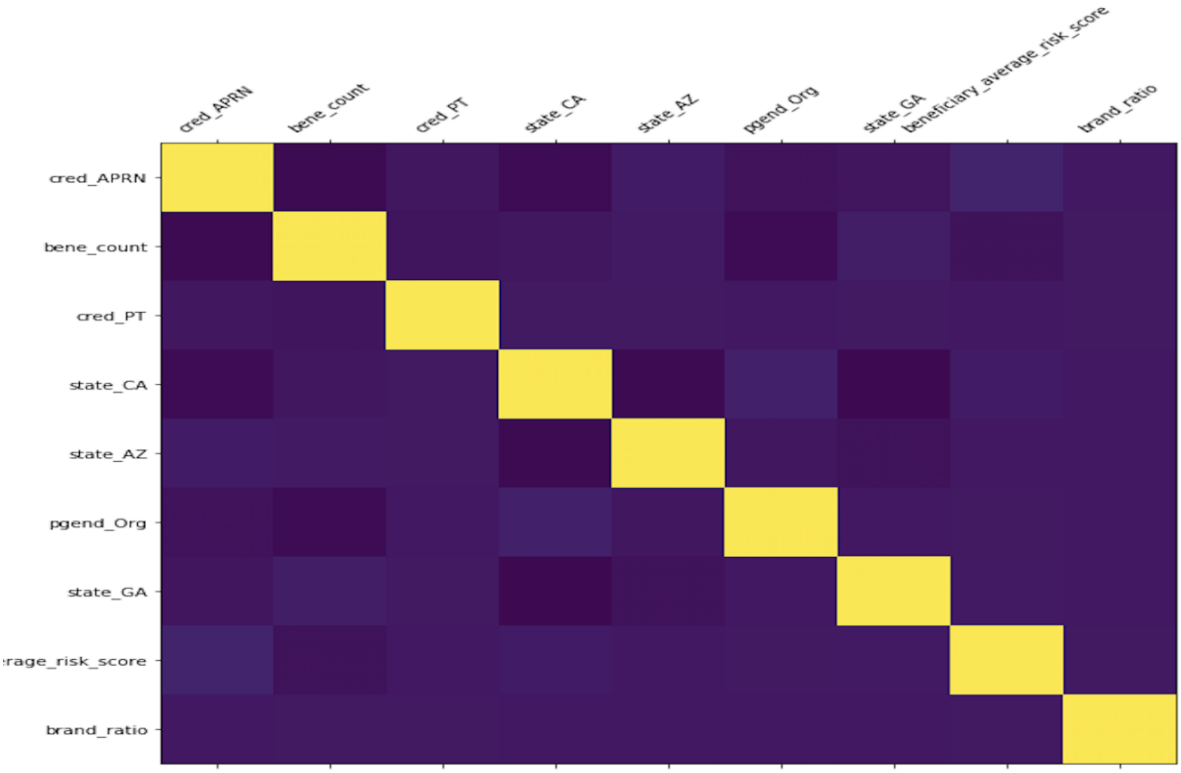


Fig-12: Correlation matrix of the Ada-boost algorithm.

Finally, we implemented an AdaBoost model. With AdaBoost, we observed similar feature importance to our decision tree models, but we obtained a higher accuracy of 55%. AdaBoost works on the principle of boosting where it creates a strong classifier using a number of weak classifiers, which is the reason we were able to obtain a higher accuracy using that model.

How exactly would the model add business value to the problem?

Knowing the various attributes that serve as predictors of medication prescription will be useful for pharmaceutical companies to market their medicines. Using predictive modeling can help companies that are looking to increase sales and help them understand the areas to target their marketing campaigns.
For example, using this model can allow companies to target doctors who they believe will be particularly active prescribers. The model can also assist in inventory management and ensure a low rate of stockouts. The models can also potentially uncover biases behind medical prescription - shining light on when medicines are over or under prescribed could help improve medical care as a whole. On the patient side, the models will be useful to help guide patients during their treatment process. They can be guided to a doctor whose care plan and medicine prescription is consistent with their preferences. The patients often have strong ideas about the use of medications. Many patients try to compare similar drugs by different companies based on various factors like price, composition, make, company loyalty etc.

Why and how the model "solve" the business problem?

Looking at the business problem from the perspective of a generic pharmaceutical company, the results of these models would allow a company to predict proclivity of doctors to prescribe branded vs generic drugs. The initial data can be fetched from npi repository (like betterdoctor.com) and Insurance company can provide

claims data for doctors for which could be compensated. From the analysis of the best performing model, we can determine which quartile has the highest ratio of branded drugs. As from a Generic pharmaceutical company's point of view this model helps them classify doctors into quartile signifying their proclivity towards Branded drugs. Generic drugs costs less than branded since it saves costs on advertisements, margins. In order for them to flourish they would need targeted advisement with high chances of results.Predicting the target variable shall help the company identify certain areas to market their drug. For example if the doctors of a certain speciality, in a particular state, or with a specific characteristic have a lesser inclination towards branded drugs, the company can devise marketing strategies to impact these doctors. This could help them channel their funds to specific areas thus providing chances of better success.

Discuss the type of evaluation metric that should be used to choose the best algorithm. How does this metric relate to the business problem?

Since false positives and false negatives are equally bad for our models and cost us equally, we have decided to use the most popular metric, 'accuracy' to decide which is the best amongst our different models. A higher accuracy in our model implies that the model can better predict which doctors/ organisations are inclined to prescribe branded/generic drugs. It turns out that AdaBoost gives us the highest accuracy (both training and testing) and hence, we have considered it to be the best model.

Identify the risks associated with your proposed plan and how you would mitigate them?

One of the risks associated with our plan is that we have assumed that we would have all the attributes of a doctor/ organization when they are added in our dataset. However, that might not be true every time. For example, the doctor may have just started practicing at the time he has been added to our dataset. As a result, we may not have some of the attributes associated with him. However, this risk can be mitigated by using the data from other datasets. For example, if we don't know the number of Asian people the doctor has treated, we may be able to approximate that count by considering the region where the doctor is operating and the number of Asian people living in that region. Also, we can use the data of other doctors in the same region to have a better approximation. Thus, briefly speaking, even though we may not have all the information regarding a doctor/ organization, we can make use of other datasets to approximate it such that the model is still fairly capable of predicting the target variable.

CONCLUSION

Now that we have identified the key drivers of brand-name drug prescription, we can deploy our model for future use. According to KFF, approximately 18,000 new doctors graduate medical school every year (kff.org). Our models can use their demographic and projected service data to estimate whether are likely to be a high prescriber of brand name drugs. For example, if a female doctor is opening her practice in Colorado, our model will predict that she will be a low prescriber of brand name drugs. Or, if the medical provider is a freshly minted graduate of a nurse practitioner program, we will expect them to be a low prescriber of brand name drugs. If the doctor plans on opening a very active practice in a high density neighborhood, we can assume that she will serve a high number of patients, and thus be a heavy prescriber of brand name drugs. Using our model that identifies the key features of doctor's prescription behavior, the pharmaceutical company can target doctors with effective marketing campaigns for their drugs, or try to convince them of the superior efficacy of their drugs compared to generic options.

One risk that is important to keep in mind is related to the bias of concept drift. Since this data came from 2013, and the incentives governing the medical insurance field change due to legislation, particularly the introduction of Obamacare, our models likely need to be re-trained on more current datasets as they become available. However, core behaviors can certainly be extracted from our dataset. To mitigate this risk, yearly updates of our models are crucial. Every year, CMS will release a year's worth of prescription data. Our models should be re-calibrated using the new data to detect if there have been any changes in prescription behavior that past years did not have.

One of the risks associated with our plan is that we have assumed that we would have all the attributes of a doctor/ organization when they are added in our dataset. However, that might not be true every time. For example, the doctor may have just started practicing at the time he has been added to our dataset. As a result, we may not have some of the attributes associated with him. However, this risk can be mitigated by using the data from other datasets. For example, if we don't know the number of Asian people the doctor has treated, we may be able to approximate that count by considering the region where the doctor is operating and the number of Asian people living in that region. Also, we can use the data of other doctors in the same region to have a better approximation. Thus, briefly speaking, even though we may not have all the information regarding a doctor/ organization, we can make use of other datasets to approximate it such that the model is still fairly capable of predicting the target variable.

When working with healthcare companies, there are always ethical considerations to take into account. In our case, increasing the amount of brand-name drugs that are prescribed could end up hurting healthcare consumers as a whole. Even though most Americans have health insurance, and so do not bear a significant out of pocket cost when purchasing drugs, there are citizens who do not have healthcare coverage and so will have to pay more to purchase the brand-name drugs. Yet, it is the consumer's obligation to educate themselves about generic equivalents and talk to their doctors about their options. Hopefully, all consumers will make educated decisions that work for their specific circumstances.

BIBLIOGRAPHY

[1] Rabon, K. and Foster, J. (11/21/14). *Generic medication problems: What to do if your prescription isn't working for you.* Retrieved Nov 2017 from http://www.thedenverchannel.com/news/investigations/generic-medication-problems-what-to-do-if-your-prescription-isnt-working-for-you

[2] Brancaccio, D. (11/18/13). *When doctors prescribe brand name drugs over generics, the taxpayers foot the bill.* Retrieved November 2017 from https://www.marketplace.org/2013/11/18/health-care/when-doctors-prescribe-brand-name-drugs-over-generics-taxpayers-foot-bill

[3] Mishori, R. (7/11/11). *Some doctors insist on brand-name drugs even when cheaper generics are available.* Retrieved Dec 2017 from https://www.washingtonpost.com/national/some-doctors-insist-on-brand-name-drugs-even-when-cheaper-generics-are-available/2011/06/13/gIQAmC0L9H_story.html?utm_term=.406abecafce

[4] Ornstein, C. and Thomas, K. (8/6/2017). Take the Generic, Patients Are Told. Until They Are Not. *New York Times*. Retrieved November 2017 from https://www.nytimes.com/2017/08/06/health/prescription-drugs-brand-name-generic.html

[5] *Statistics and Facts on U.S. Physicians/Doctors*. Retrieved Dec 2017 from https://www.statista.com/topics/1244/physicians/

[6] Dingwall, N., Potts, C. and Senaratna, D. (9/13/16). *Prescription-based prediction.* Retrieved October 2017 from https://www.nytimes.com/2017/08/06/health/prescription-drugs-brand-name-generic.html

[7] The Centers for Medicare and Medicaid Services, Office of Enterprise Data and Analytics. (5/25/17). *Medicare Fee-For Service Provider Utilization and Payment Data.* Retrieved October 2017 from https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Prescriber_Methods.pdf

[8] Consumer Reports. (9/30/14). *Find out if your doctor takes payments from drug companies*. Retrieved November 2017 from https://www.consumerreports.org/cro/news/2014/09/find-out-if-your-doctor-takes-payments-from-drug-companies/index.htm

[9] Orenstein, C., Jones, R., and Tigas, M. (3/17/16). *Drug-Company Payments Mirror Doctors'*

APPENDIX 1

Group Contributions:

Project Selection and Dataset Collection: All group members surveyed at least 5 data topic ideas and submitted them to the group, along with a dataset that could support the idea. In collaboration, the group narrowed down our choices to three for a final vote, and all group members voted on the final topic chosen.

Business Problem Development: Again, this was a very collaborative process, with all group members contributing ways in which the data could be used to help solve or alleviate a business problem. Jayesh in particular kept the group focused on using the model to solve a concrete business problem that had applications for future deployment.

Data Cleaning and Preparation: Sanket began the data prep with an initial framework for loading and setting up the dataframe. Ben and Jayesh took the primary lead in the data cleanup, with coding help from all team members.
Modeling and Evaluation: All team members were assigned a modeling task. Titash implemented Decision Trees and Random Forest, Sanket implemented Logistic Regression and Ada Boost, Ben handled three types of Linear Regression and Jayesh combined all models for comparison. All team members were responsible for recording accuracy, importance of feature variables, and generating any relevant plots and visiualizations for their respective models. Once collected, the team collectively compared the results of the different models.
Deployment: Group collaborative discussion on how the model would be used in practice, as well as what challenges it might face in deployment. The group discussed ethical considerations as well. Ben worked on editing and writing the data cleanup in the project report. Titash combined the results of all the models and worked on editing the project report, adding the graphs and presenting the results.

APPENDIX 2

Imputation of Missing Values in Drug Claim Count Fields

There are three fields that relate to the counts of drug claim counts: brand claim count, generic claim count, and other claim count. Brand claim count refers to all claims submitted to Medicare Part D for brand name drugs, and the same for generic claim count for generic drugs. Other claim count refers to number of claims for drugs that did not fall into either brand or generic categories. The documentation on these kinds of drugs on the CMS website does not really explain what they are.

The other claim count field is generally zero or blank, as there were not many cases where other drugs were prescribed. However, due to the CMS rules protecting patient privacy, any case where one of the three types of claim counts is between 1 and 11 claims, the values in two of the three fields are suppressed. Since the other drugs are rarely prescribed, this caused a lot of the data to be suppressed in the key brand and generic claim count fields that we needed to create our target variable ratio.

CMS did provide an accompanying column for each of the three claim count fields that indicated whether the data that was suppressed was for a value over 10 claims or under 10 claims. This helped us fill in the missing values using the following methodology:

Any case where the suppressed value was indicated to be less than 10 claims, we estimated that it was 1 claim. The choice of 1 was our best option since the other claim count field was most often suppressed due to low counts, and other claims are generally pretty rare. Once we estimated all cases where the suppression was caused with low values, we could estimate the other field that was suppressed (but which had a claim count higher than 10, and was only suppressed due to the other field being suppressed) by taking the difference from the total claim count, the field that was not suppressed, and the field that had the one count estimation.

To analyze if this estimation methodology made a big difference when modeling, we established two data sets for use in modeling. One has only rows where no values were suppressed in the count fields, and the other has all relevant rows, with estimations to fill in missing values. The results of the comparison between these two data sets can be found later on in the main body of the report.