PharmaCo Hackathon Solution

Alliyah, Tenbite , Belinda, Juan, Yuping

Hult International Business School

Executive Summary

Business Problem

- How can PhramaCo increase profitability in its industry?

Diabetes Dataset

- Datasets include multiple variables for patients taking different medications. It also includes information on their medical diagnosis, facilities they are in as well as the different reasons to visit the hospital.
- The medications mentioned in the dataset are used in order to treat patients with diabetes.
- Missing values in the data were filled in with mode in order to create a filled dataset to make further analysis.

Assumptions During Data Massaging

- The team made assumptions that some medications could be more effective for patients to take to treat diabetes compared to others.
- The data can also tell us information on the demographic data that explains who the patients are and their behaviors explained in the data.

Key Insights from Variables

- As the race bin increases, the diabetic medications also increase showing an increasing trend for African-Americans, Asians, Hispanics & Other uptake of diabetic medications.
- The payer_code displays that the self-pay group, insured individuals & unknown group take diabetic medications the most.
- Lastly, there is a trend that younger people take more medications than the older group.

Recommendation

- Race: Target African Americans & others
- Payer Code: Partner with main insurance companies
- Age: Focus on younger age groups
- Medication: focus on the platform for pharmacies that provide insulin medications as this is in high demand.

Business Problem

As PharamCo seeks to continue its role as a leader in the pharmaceutical industry, they are faced with finding ways to increase its profitability. After researching the different kinds of diseases that are occurring it is determined that focusing on the diabetes industry could have room for growth and profitability.

Research projected that 39 million Americans, which constitutes 13.9 % of the population, will have diabetes by the year 2030 (Lin J, 2018). This population will be continuously looking for a variety of drug solutions. In total, there are 159.1 million diabetic drugs being prescribed showing a huge opportunity for Pharmaco to provide a solution to patients suffering from diabetes.

Diabetes Dataset

The provided dataset included information on patients' admission based on their hospital visits with 101,766 observations and 51 variables. Variables highlighted in the data included specific patient codes, prescriptions, diagnosis type, demographics, A1C results, readmission status, change in medication information, and more. The descriptive statistics of the data show that the mean of patient admission type is 'urgent', and on average patients are either 'transferred to SNF or ICF' in discharge disposition. On average the time patients spent in the hospital is '4.4' days and 50% of the sample had 44 lab procedures during the period. Moreover, patients in our data were prescribed 16 medications on average.

In the data cleaning process, we began to analyze missing values that were identified as '?' and filled the values with mode, '0', or unknown where appropriate. The following variables that were filled with the mode include race ("Caucasian"), gender ("Female"), weight ("0"), payer code ("MC"), and medical specialty ("unknown"). The purpose of filling missing values with 0 or unknown is to transform these variables into bins during the data massaging process. The next step in our cleaning process included creating bins in order to organize character variables. For example in the age column grouping was done in (1, 2, and 3) which specifies young, middle-aged, and old age. Another example of the bin we created included an admission type which was organized into <12 hours of treatment, >12 hours of treatment, newborn, and unknown.

Assumptions During Data Massaging

After creating bins, the data was made into charts including histograms to determine which bin was having a significant impact on each variable. Looking at correlations also helped in identifying which variables are impacting each other in order to implement regression models to further analyze the data.

Assumptions the team made while cleaning the data were that older patients would be the ones that could have critical cases of diabetes. Also, there are medications in the data that could be more effective than others. Additionally, we believed that all the independent variables in the data could have an impact on whether the patient would be discharged after 30 days or before based on their diagnosis code. Identifying these variables determine how most diabetic individuals are effectively treated. See appendix C for key insights into the variables.

Correlations

Insights display that there is a correlation between number_medications and number_lab_procedures by 0.39. This means that patients who are prescribed medications would have had several lab procedures prior. An additional correlation of 0.26 between number_emergency and number_inpatient means that patients who have visited the hospital for emergency reasons are registered as an inpatient.

Predicting Factors that Impact Diabetic Diagnosis

Our regression model, according to appendix B, reveals that race, discharge disposition id, payer_code, age, admission referral, and medical specialty are all significant in predicting people who take diabetic medications. Insights discovered from this predictive model reveals the following:

1. As the race bin increased, the diabetic medications also increases by 0.04 units. Even though the number is low, it shows an increasing trend for African-Americans, Asians, Hispanics & Other uptake of diabetic medications.

2. Within each bin for payer_code displays that 83 % of the self-pay group, 80 % of insured individuals & 73 % of the unknown group take diabetic medications.

3. Moreover, within each age group, there is a trend that younger people take more medications than the older group (82 % vs. 76 %).

Recommendations

1. **Race:** Based on our predictive model, we recommend that PharmaCo target African Americans, Asians, Hispanics & others as they are more likely in the future to use more diabetic medications.

2. **Payer Code:** PharmaCo should focus on partnering with main insurance companies to locate the most affordable medications for diabetics. This will be done by PharmaCo selling their services of locating the medications for insurance companies to lower cost & time. This again will increase PharmaCo's profits.

3. **Age:** Focus more on younger age groups between the age of 0 - 30 years, because the possibility of a young person having diabetes today is higher than the past. Also, it is easier for the younger group to accept using tokens from PharmaCo as they are more tech-advanced in society.

4. **Medications:** Focusing on insulin, metformin and glipizide will be more profitable as these are the medications that are being used constantly by diabetic patients.

References

Lin, J., Thompson, T.J., Cheng, Y.J. *et al.* Projection of the future diabetes burden in the United States through 2060. *Popul Health Metrics* 16, 9 (2018). https://doi.org/10.1186/s12963-018-0166-4

Kane, S. P. (n.d.). *ClinCalc DrugStats Database*. CLINCALC DrugStats database. Retrieved January 25, 2022, from https://clincalc.com/DrugStats/Default.aspx

Appendices

Appendix A

Dictionary for variables with bins:

Race:

1. Caucasian
2. African-American
3. Other

Admission type ID:

1. < 12 hours treatment
2. > 12 hours treatment
3. Newborn
4. Unknown

Discharge disposition id:

1. Discharged Home

People who can go back home or only need daily care out of hospital
(1, 3, 4, 6, 8, 24, 27, 30)

2. Hospital

People who still need to stay at the hospital
(2, 5, 9, 10, 12, 15, 16, 17, 22, 23, 28, 29)

3. Expired

People who expired or near expired
(11, 13, 14, 19, 20, 21)

4. Non-conformist

People who don't follow medical advice / group of people with unknown discharge disposition ID
7, 18, 25, 26

Payer_code:

1. Unknown
2. Other insurance
3. Medicare/Medicaid
4. Self-pay

Age:

1. Young (0-30 yrs)
2. Middle (30-60 yrs)
3. Old (60-100 yrs)

Admission referral source ID:

1. Unknown
2. Babies
3. From low risk medical institutions
4. From high risk medical institutions
5. Other

Medical specialty:

1. Popular medical specialty
2. Unpopular medical specialty
3. Unknown

Note that popular medical specialties are specialties with more than 1000 physicians in that group, while unpopular specialities are less than 1000 physicians.

Bin 1 includes:

- Emergency/Trauma (1 variable)
- Family/General Practice (1 variable)
- Cardiology (1 variable)
- InternalMedicine (1 variable)
- Surgery Procedures (13 variables)
- Pediatrics (11 variables)
- Orthopedics (2 variables)

Bin 2 includes:

- AllergyandImmunology, Anesthesiology, DCPTEAM, Dentistry, Dermatology, Endocrinology, Endocrinology-Metabolism, Gastroenterology, Gynecology, Hematology, Hematology/Oncology, Hospitalist, InfectiousDiseases, Nephrology, Neurology, Neurophysiology, Obsterics&Gynecology-GynecologicOnco, Obstetrics, ObstetricsandGynecology, Oncology, Ophthalmology, Osteopath, Otolaryngology, OutreachServices, Pathology, PhysicalMedicineandRehabilitation, PhysicianNotFound, Podiatry, Proctology, Psychiatry, Psychiatry-Addictive, Psychiatry-Child/Adolescent, Psychology, Pulmonology, Resident, Rheumatology, Speech, SportsMedicine, Urology

Bin 3 includes: Unknown (1 variable)

Appendix B

Regression model

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           1.541522   0.079784  19.321  < 2e-16 ***
race_bin              0.037423   0.014487   2.583 0.009790 **
admission_type_bin    0.009664   0.009074   1.065 0.286829
dis_dis_bin          -0.143377   0.009401 -15.252  < 2e-16 ***
payer_code_bin        0.190342   0.008410  22.632  < 2e-16 ***
age_bin              -0.087136   0.014631  -5.956 2.59e-09 ***
admission_referral_bin -0.056537   0.015418  -3.667 0.000245 ***
medical_specialty_bin  -0.069114   0.008317  -8.310  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 109752  on 101765  degrees of freedom
Residual deviance: 108738  on 101758  degrees of freedom
AIC: 108754

Number of Fisher Scoring iterations: 4

>
```

PREDICTIVE MODELING

Diabetes Medication = 0.04 race_bin + 0.01 admin_type_bin - 0.14 dis_dis_bin + 0.19 payer_code_bin - 0.09 age_bin - 0.06 admin_ref_bin - 0.07 med_spec_bin

Note: The only insignificant variable Admission Type Bin

Appendix C

Key Insights from Variables

The following highlights the largest frequency of each variable below:

Medical Speciality
Unknown - 49,949
Common_Med -37,059
Unknown - 14, 758

Adm_Refferal
High Risk Facilities (60,691)
Low-Risk Facilities (37,637)

Payer_Code
Unknown - 40,256
MC -32,439
Other Insurance 24,064

Age
Old (60-100) - 68,541
Middle age (30-60) - 30,716

Race
Caucasian (78,372)
African American (19,210)

Admission Type
<12 hrs Treatment 72,491
>12 hrs Treatment 18,869

Discharge_Disposition
Go Home Af-Discharge (88,066)

Appendix D

NUMBER OF PEOPLE PER AGE GROUPS BY DIABETES MEDICATION

```
> xtabs(~ diabetesMed_dummy + age_bin,data = df_diab_data)
                 age_bin
diabetesMed_dummy     1     2     3
                0   461  7061 15881
                1  2048 23655 52660
> |
```

	Young	Middle-Aged	Old
No Diabetes Medication Taken	**18.4%**	**23.0%**	**23.2%**
Diabetes Medication Taken	**81.6%**	**77.0%**	**76.8%**

Appendix E

RACE GROUP TAKING DIABETES MEDICATION

```
> xtabs(~ diabetesMed_dummy + race_bin,data = df_diab_data)
                  race_bin
diabetesMed_dummy     1     2     3
                0 18051  4412   940
                1 60321 14798  3244
>
```

	Caucasian	African-American	Hispanics & Others
No Diabetes Medication Taken	23.0%	23.0%	22.5%
Diabetes Medication Taken	76.97%	77.03%	77.53%

Appendix F

- **Projected market value from African-Americans & Hispanics by 2030 is $6.67B**