Oncology Research Platform
MS Analytics and Deloitte Practicum – Summer 2015
Predictive Models Write-up

We used two predictive analytics tools – logistic regression and a classification tree. The R script used for this analysis (PracticumPredictive.R) is also included. The cleaned data derived from the SEER database used for this script (BreastCancer2010-2012.csv) is also included.

Logistic Regression:

From the exploratory analysis, we found that the variables that had a significant relationship with survival were age at diagnosis, race, marital status, stage, grade of tumor, subtype of tumor, size of tumor and number of positive nodes. A full list of these variables, their type (categorical/continuous) and the values they can take can be found in appendix [1].

The subtype of a patient (HER status) recorded in the SEER dataset starting from 2010 through to the end of the study in 2012. So we looked at only the year 2010 so that there would be a substantial amount of time to actually determine if the patient survived or died. We first attempted to create a predictive model for survival from the 2010 dataset and found that survival for stages 1.2 and 3 was extremely high (96%). However for stage 4, survival was much lower (58%). So we decided to focus on prediction for patients diagnose in stage 4. We did a stepwise logistic regression with the Event as the dependent variable (Alive or dead because of other causes = 0, Dead because of cancer = 1) and the above-mentioned variables as predictors. Using a base model with only the intercept as a predictor and a full model with the all the above variables as predictors (and their second order interaction terms), we ran a stepwise regression in both directions, forward and backward. The model that was selected after this stepwise regression included the variables subtype of tumor, grade of tumor, size of tumor, age at diagnosis and race. This model had the lowest AIC. The AIC or the Akaike Information Criterion measures the information loss for a given model relative to the other models. The full model is included in appendix [3]. Most of the variables in this model also were statistically significant (at a 0.05 level).

To see how well our model would work on an independent data set, we used k-fold cross validation. We used k=10 to divide the model into 10 parts, using 9 parts as the training set and the 10^{th} part as the testing set. Repeating this process 10 times with each of the 10 parts as the testing set, we calculated the average area under the ROC for the 10 "folds" which came out to be 0.74gma. The ROC or the receiver operating curve measures the performance of a model by plotting the true positive rate against the false positive rate of the model. The area under the ROC quantifies the probability that the selected model will correctly classify a randomly selected observation as survived or dead.

Classification Tree:

The second modeling technique we used was a classification tree. This tree was also built using the 2010 stage 4 data. A classification tree uses a set of if-then rules to

classify each observation into a binary outcome. We used the "rpart" package in R to build this decision tree. This package uses the CART (Classification and Regression Trees) algorithm that uses a measure called the gini impurity to decide the splitting for a decision tree. The full classification tree can be found in the appendix [4]. This tree has one node – splitting on the variable subtype. If the patient belongs to subtype 1, 2 or 3 – they have a 0.67 probability of survival and if the patient belongs to subtype 4 or 5, they have a 0.35 probability of survival. This simple tree with one node has a classification error (also known as generalization error) of 0.33.

It is possible to get better results by including more of the variables in the SEER dataset, using different modeling techniques such as random forests or recursive partitioning decision trees. However for the purposes of the demonstration for this research platform, both these models are sufficient.

Limitations of the dataset:

The data we used has certain limitations that may affect our results. This data is only for the years 2010-2012. While the SEER dataset has data starting from 1973., we only used data from 2010 since some important variables such as subtype of tumor were recorded only from 2010 onwards. Another variable – size of tumor was recorded since 2007. Additionally, a lot of research has been done in recent years in breast cancer that makes the newer data more relevant.

Another limitation is that the SEER dataset is only collected for 10 regions across the US. So it is possible that if this data was collected for more regions, other variables may become significant or show a relationship with survival.

Oncology Research Platform
MS Analytics and Deloitte Practicum – Summer 2015
Predictive Models Write-up

Appendix:

Appendix 1 – Variables included for predictive modeling

Variable Name	**Variable name in SEER data**	**Description**	**Type**	**Possible Values**
EVENT	SEER cause-specific death classification	Whether the patient survived or not.	Binary	0 = Patient alive or dead of other causes; 1 = Patient dead of cancer
Race	Race recode	Race of the patient	Categorical	1 = White; 2 = Black; 3 = American Indian/Alaska Native; 4 = Asian/Pacific Islander; 7 = Other; 8 = Unknown
Age_Diagnosis	Age at diagnosis	Age of patient at diagnosis	Continuous	0-130 (years); 999=unknown
Marital_Status	Marital Status at dx	Marital status of patient at diagnosis	Categorical	1 = Single(never married); 2 = Married; 3 = Separated; 4 = Divorced; 5 = Widowed; 6 = Unmarried or domestic partner; 9 = Unknown
Grade	Grade	Grade of tumor	Categorical	1 = well differentiated; 2 = moderately differentiated; 3 = poorly differentiated; 4 = undifferentiated; 5 = T-cell; 6 = B-cell; 7 = Null cell; 8 = N – K cell; 9 = undetermined
Subtype	Breast subtype	Subtype of	Categorical	1 = HER2+/HR+; 2 =

		tumor		HER2+/HR-; 3 = HER2-/HR+; 4 = Triple Negative; 5 = Unknown
Tumor Size	CS Tumor Size	Size of Tumor (in mm)	Continuous	000-989 (exact size in mm)

Appendix 2 - Stepwise logistic regression full model:

```
Call:
glm(formula = EVENT ~ Subtype + Grade + Age_Diagnosis + Race +
    Tumor_Size, family = binomial(link = "logit"), data = data2010stage4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0873  -0.9037  -0.6305   1.0400   2.2952

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.855204   0.456720  -6.252 4.06e-10 ***
Subtype2        0.080131   0.290962   0.275 0.783009
Subtype3       -0.010783   0.209627  -0.051 0.958975
Subtype4        1.334156   0.253414   5.265 1.40e-07 ***
Subtype5        1.026712   0.254890   4.028 5.62e-05 ***
Grade2         -0.347601   0.314030  -1.107 0.268335
Grade3          0.293353   0.310091   0.946 0.344137
Grade4         -0.470559   1.201455  -0.392 0.695311
Grade9          0.899940   0.323112   2.785 0.005349 **
Age_Diagnosis   0.029092   0.004968   5.856 4.74e-09 ***
Race2           0.659438   0.176500   3.736 0.000187 ***
Race3          -1.072981   0.634876  -1.690 0.091016 .
Race4          -0.019154   0.259289  -0.074 0.941113
Race9         -13.232713 342.648712  -0.039 0.969194
Tumor_Size      0.002042   0.001127   1.811 0.070169 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1541.3  on 1130  degrees of freedom
Residual deviance: 1336.9  on 1116  degrees of freedom
AIC: 1366.9

Number of Fisher Scoring iterations: 12
```

Appendix 3 – Classification Tree (using CART algorithm):