**DATA622: Homework3: Essay**

by Glen Davis

**Part 1: Support Vector Machine (SVM) Models vs. Tree Models in My Analysis:**

For Homework 3, I trained radial basis and linear SVM models on [the same dataset of Web sites labeled either Phishing or Legitimate that I used for Homework 2](https://www.kaggle.com/datasets/danielfernandon/web-page-phishing-dataset). As a reminder, in addition to the binary response variable, there were 19 integer predictor variables. Most of them represented counts of specific punctuation characters within the Web sites’ urls, like slashes or dots, and there was also a count of the total number of characters within the url, as well as a count of redirects within the url. Many of the predictors demonstrated near-zero variance and were removed from consideration for all models in order to eliminate noise.

Because SVM models can be slow to train on large datasets, and my dataset had a somewhat large number of observations, I reduced the percentage of data used to train the models from the typical 70% to a more manageable 10%. This reduction seemed a little extreme at first, but I made several attempts to use larger percentages of the data to train the models, and the computation costs only became reasonable at this figure. The gain in speed when building the models was worth it, and I confirmed later that the models still performed very well despite using only a fraction of the data on which the tree models I built in Homework 2 were trained.

The radial basis SVM model was more accurate than the linear SVM model, and it had a better balance between precision and recall. Although both SVM models were more accurate than the two Decision Tree models from Homework 2, neither was as accurate as the Random Forest model from it. The Random Forest model also had the best balance between precision and recall overall, as well as the best recall period, but it is worth noting that the linear SVM model beat it in precision. Had the SVM models been able to utilize more data during training, I do believe there's a decent chance they could have outperformed the Random Forest model though. Training the SVM models seemed slower than it should have been, even accounting for my somewhat large dataset, and I need to confirm whether the sluggishness was due to the limitations of the libraries I was using or some other factor(s).

**Part 2: SVM Models vs. Tree Models in the Academic Literature:**

In “Decision Tree Ensembles to Predict Coronavirus Disease 2019 Infection: A Comparative Study,” I found the discussion of the special considerations required when classifying imbalanced data using tree models enlightening. I have worked with imbalanced data before, and I am particularly interested in adopting Area Under the Precision-Recall Curve (AUPRC) as a good performance measure for imbalanced data classification tasks in the future. I also found it interesting that using Synthetic Minority Oversampling Technique (SMOTE) to correct for the class imbalances worsened the performance of the ensemble tree models here. I’ve seen many people on the Internet express opinions that SMOTE either doesn’t help or actively harms a lot of data analysis, but this is the first paper I’ve read that documents its negative effects on an analysis explicitly.

In “A Novel Approach to Predict COVID-19 Using Support Vector Machine,” I found the performance of the SVM model using a pretty small number of observations and features remarkable despite the note that the K-Nearest Neighbors model actually performed best using the least data. Nothing else really stood out about this article, other than the authors’ thoughtfulness during feature selection. I always appreciate a limited number of features.

Natural Language Processing (NLP) is not necessarily my area of expertise, but it is something I’ve taken a special interest in throughout this program, and I have been trying to build up my NLP toolkit this semester in particular. So I chose three articles that compared the performance of tree models and SVM models in the context of NLP, and here are my thoughts on them:

* [Comparison of Naïve Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification](chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https:/www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/5_2_05_Pranckevicius.pdf)
  + Logistic regression outperformed all other models in this multiclass classification experiment, including the tree models and the SVM models, but its predictive abilities weren’t stable across all classes, meaning it was better at predicting some classes than others. The rest of the models’ predictive abilities were more uniform across classes.
  + Decision Tree models were the worst performers.
  + Unigrams alone were sufficient—adding bigrams and trigrams didn’t improve performance in a statistically significant manner.
  + After a certain point, i.e. 5,000 observations per class, adding more observations per class didn’t improve performance in a statistically significant manner either.
  + One drawback of this article is that because the analysis was done in Apache Spark, the only kernel function the authors could use for their SVM models was linear.
* [Question Classification Using Support Vector Machines](chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http:/www.kernel-machines.org/papers/upload_29525_p31189-zhang.pdf)
  + The authors make a good point in discussing the differences between question classification and typical document classification: stopwords should not be removed in the case of question classification. I have actually found this to be true in many NLP situations, as it leads to reduced model performance. The same can be said for stemming and lemmatization much of the time. Context is key.
  + The authors make the same conclusion that the previous article did: unigrams are sufficient for analysis, and adding ngrams of any length did not improve performance.
  + No kernel function performed better for the SVM models than the linear one, interestingly. I have not found this to be the case in my analysis, so this seems specific to question classification.
  + SVM models outperformed all others [Nearest Neighbors, Naïve Bayes, Decision Trees, and Sparse Network of Winnows (SNoW), which was a new one to me].
* [Random Forest and Support Vector Machine Based Hybrid Approach to Sentiment Analysis](chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https:/pdf.sciencedirectassets.com/280203/1-s2.0-S1877050918X00039/1-s2.0-S1877050918301625/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEFoaCXVzLWVhc3QtMSJGMEQCIDv4EmB7dzRYPPGaTnW5WBMNSBBNC412ykCK6c4curU%2BAiBL2bC0SYXRZzhACwNkPett%2FgAFCSbmLqx0HPGlQvYv5Sq8BQiy%2F%2F%2F%2F%2F%2F%2F%2F%2F%2F8BEAUaDDA1OTAwMzU0Njg2NSIM%2Fa6EZBO2FPNze6ynKpAF2bKz4WXrgQiW2yU9Gl%2BxFdISI7x4ztPZqk2Gcjg0y4preZM6xFRo%2FjJsfbALov5fKCT3Vu%2BoYzBRCXh6uAq4GyBtCnr0ESbmvRHt4Bb0eYubDmoZjm55DSpNILUrR4a%2FXpvYcxjUT0PW8%2B9okHTgXCcLml5kTjK2ITN0vrhBeWnkS6YT5ia2mBMaJtLmNRKA91FMwrsbl24gevHtdquESOf4Kw%2FyLbo5%2BH%2Fwmu5eQAUPrhUNf%2FnOdmnN1uQK1YC9GnkvK9pHVgzixHUQzPD%2Bd7Sh9zSkDCxn%2B2gJGldt6F1Bg0PRQPe4tkCm2q1HCX14tM%2FRT%2FwSRxgoOSbnE4CpybOp%2FyeRqiMjhe9gLf1ohfC%2BpdGHfsMtK1qJR8eD2g61B6%2FPRqL5awm34dvdeVIMVOS6TT8Mb2ODBMX5CvgvajtVjiHgmE8imn%2FpQMm99iUNhsq7duGVC%2FaR9fuh2ZLH3x4aLqIY%2BmZThT%2BkarBK5TymdmCkjVTVBbBX5QhQfwnbzQiFcbP03s79bMxeetlUYsmj9Fof3I7mrfoqVDaj2xn%2FWMfqfng3K%2FY78x4TH19VpFAAtkLySLzTQTKiseUmjVlhBVamJAM19CQJfjP%2FtUGickoRRBqLKWNqoRETqC%2B9BDJlfgGNyi1byIiNdqyon3C2WrjT4ThQO7fBTUJFqMQRlF278UopQ3GxzT9IMdPX784Xv1paA2rRxcbKRuw6WDni%2Fa8eEolL9oxhIB%2BOoLEwmLnyOXoOKYZyVA6wtDyDqcdL60D%2BHZ7qUXW%2F66hYrwQAP6jbmVmcpjn4JGVhhSXT0nmdzFhLwcsPAhsJRabSWnvBz8rVWNlcUbp5uCLDdjnRe1v9ZrugcPaQ9fdxfykw49vgsQY6sgH0w6L7w2bEx8zsJsCus55Kurmb7ZEmpQUOqNgZwtLuesUN30M4U6nAp%2Fp02hlDJ%2FS8KIi%2BrbbHzylKUx2IsbOsVPYDB4PVjZqrkxm5rhfXAaYxdJZuB1fTQrcZlgHpGETtQBRIqikPlG7Nm3484GHvEBTfqYHzEfQ9O7tyBw1ZrDNiWnASOi1FyOYiYmxrejmtxKGdin42g6BqBtzogA7mZQQ9ldkwr%2FRGTmy3dLol8gWb&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20240506T015212Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTY3X6524PI%2F20240506%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=3aad554d83a841df59d69b89a8b0ae13dd89d912f22dbb45e989c6d6713b47cf&hash=ce5c839901fcc0c1fb3443a22223f3b4069428e9bdd96bf3e62210d930909a0f&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S1877050918301625&tid=spdf-d82deb1c-9ebb-4e16-bfc6-a12538c7ba91&sid=e22adfca3607844b2b486826)
  + I found this to be the least useful of the three articles. It spends a lot of time defining Random Forest and SVM models and very little time accurately explaining how they combined the two in their hybrid approach to sentiment analysis. As best as I can tell, they just averaged the “positive” vs. “negative” scores from both models and called the resulting classifications a hybrid model. But this hybrid method does perform better than either of its individual components, so I guess it was a worthwhile experiment.