**Module 1 Discussion:**

Nearest neighbor models are popular in classification problems. After working through this module, identify and discuss a situation (in business or just in life) that you believe that this type of model would be applicable. Make an initial post connecting this week’s learning to the prompt.

A decision problem with instances or examples of training data that are deemed important or required by the model is an instance-based learning model. In order to discover the best match and generate a forecast, such algorithms often build up a database of example data and compare new data to the database using a similarity measure. As a result, instance-based approaches are also known as winner-take-all and memory-based learning methods. The representation of the stored instances and the similarity measures used between instances are both highlighted. KNN is frequently used in decision-making models, recommendation systems, and image recognition technology. Although, in the actual world, the recommendation system employs more advanced algorithms.

KNN isn't ideal for high-dimensional data, but it's a great starting point for systems. Netflix, Amazon, YouTube, and a slew of other companies provide individualized recommendations to their customers. It's the algorithm that firms like Netflix and Amazon use to suggest movies and books to watch or buy. They'll use KNN on a data set compiled from the movies you've seen or the books you've purchased on their website. These organizations will then take your customer information and compare it to that of other customers who have seen comparable movies or purchased similar books. Using KNN, this data point will be classed as a specific profile based on their history. The movies and books that are recommended will then be determined by how that data point is classified by the algorithm. KNN can look for documents that are semantically similar. Each document is treated as if it were a vector. If two documents are next to each other, it suggests they have the same themes. KNN is a powerful tool for finding outliers. Credit card fraud detection is one such example.

**Reference:**

[1) Vaibhav Jayaswal. (Aug 25, 2020). K-Nearest Neighbors (KNN) algorithm. *Towardsdatascience Medium Article*. Retrieved from https://towardsdatascience.com/k-nearest-neighbors-knn-algorithm-23832490e3f4

**Module 2 Discussion:**

This scene from Moneyball (Links to an external site.) highlights how statistics can change the strategy used in putting together a winning team. In this case, the target variable was changed from wins to runs, and that changed how management acquired new players. How does that mindset and linear regression inform that this is a good strategy?

Moneyball is a film that depicts the tale of the Oakland Athletics. Billy Beane used an analytical, sabermetric approach to creating a competitive team. College baseball players are a better candidate for higher offensive Moneyball statistics. Some abilities are devalued while others are inflated. In baseball, money is a major factor in player selection. There are two hypotheses being utilized to restrict the selection process. The Moneyball hypothesis pays no attention to the athlete's body or athletic abilities. Billy Beane decided to pick position players/hitters based on statistics. The Moneyball idea is unaffected by the athlete's physical structure or the tools he or she possesses. Billy Beane (the idea for Moneyball) recommended launching position players/hitters based on certain statistics. On-base percentage (OBP) and slugging percentage were the two statistics. On-base plus slugging is a new statistic (OPS was formed as a result of combination of these two stats).

Billy Beane was motivated to develop this hypothesis by a sabermetrician named Bill James. Beane believed in selecting experienced college players over inexperienced high school amateurs. Billy Beane absorbed LeBron James' mentality when it comes to lining up his players. From a coach's perspective, part of a hitter's duty is to hit homeruns, singles, doubles, get on base, drive in runs, and steal bases. A hitter's job, according to James, is to generate runs.

A formula that allows one to establish created runs:

(Hits + Walks) x Total Bases  
At-bats + Walks

Billy Beane's produced runs algorithm is 90% accurate and delivers a sum of the team's real scored runs within 5% of the time. The only way to score runs is to get on base, and because walks are such a big element of the calculation, the on-base percentage should be closely monitored.

It works off the simple formula:  
(A x B)/ C  
The A variable adjusts the “on-base” aspect of baseball.

A = hits + walks + hit batsmen – caught stealing – ground into double play (H + W + HBP – CS – GIDP)  
The B variable takes into account the advancement of the player.  
B = total bases plus .35times hit batsmen and non-intentional walks, alsop.70 times stolen bases, sacrifice hits, and flies (TB + .26(TBB – IBB + HBP) + .52(SB + SH + SF)

The C variable holds for opportunity.

C = at-bats + total walks + sacrifice hits and flies + hit batsmen (AB + TBB + SF + HBP)

Billy Beane's 'Moneyball' was a success for the Oakland Athletics not because of data analytics, but because he restructured the organization to enable it to realize its potential. This study shows how a low-budget organization can generate high-quality baseball players. It discusses the benefits of incorporating data science into an organization's DNA.

**Reference:**

[1] Ehren Wassermann, Daniel R. Czech, Matthew J. Wilson & A Barry Joyner. (n.d.). An Examination of the Moneyball Theory: A Baseball Statistical Analysis. *The Sport Journal*. Retrieved from https://thesportjournal.org/article/an-examination-of-the-moneyball-theory-a-baseball-statistical-analysis/

**Module 3 Discussion:**

The article below talks about how government agencies are incorporating binary models to handle detained families. This is a perfect example of how classification models have real-life implications on people. https://www.washingtonpost.com/immigration/immigrants-choice-separated-children-jail/2020/10/23/d7cf18b2-1545-11eb-82af-864652063d61\_story.html

**Discuss:**

* Do you think this is a valid way of using a predictive model?
* What kinds of variables would you consider if you had to use this type of model?
* Is this a model where accuracy or speed is most important?

**Do you think this is a valid way of using a predictive model?**

Gee's binary-choice model can provide parents the ability to choose where their children go, whether they are released to another approved caregiver or whether they waive their children's right to release. This, in my opinion, is a valid application of a prediction model. Before filing a case to the court, it serves as a classification tool. It can also take use of data processing and mining results to provide adequate therapy for migrating children.

**Variables Consideration:**

The first variable I would consider for this model is the rationale for immigration, whether it is reasonable or implausible. The nationality of the speaker is also relevant in judging the veracity of their claims. Then I consider basic family facts, family income, and obligations, in order to get a quick image of their family and determine whether they are suffering greatly. We may also keep track of the children's age, gender, health issues, and, if possible, their preference for staying with their parents during incarceration or being separated.

**Is this a model where accuracy or speed is most important?**

This binary-choice model's accuracy should take precedence. While one goal of the approach is to increase court efficiency, it was designed to provide parents specific rights rather than leaving it up to the courts. In order to be humane, the accuracy of the model should take precedence above speed, regardless of the erroneous findings obtained. The model's output may result in a difference in the fate of the families; thus, judgments should be made with caution.

**Reference:**

[1] Miroff, N. (2020, October 23). Migrant parents could face fateful choice: Be separated from their children or stay together in jail. *The Washington Post*. https://www.washingtonpost.com/immigration/immigrants-choice-separated-children-jail/2020/10/23/d7cf18b2-1545-11eb-82af-864652063d61\_story.html

**Module 4 Discussion:**

In the following article, Instacart Market Basket Analysis (https://medium.com/kaggle-blog/instacart-market-basket-analysis-feda2700cded), the runner up in the Kaggle competition talks about analysis they did to predict order sizes for Instacart. Their main approach was to use xgboost (gradient boosting). In this week’s discussion, cover the following points:

* What problem was the user trying to solve?
* Do you think there were other models they should have used? Justify your answer.

**What problem was the user trying to solve?**

Instacart is an e-commerce platform where customers may order products from neighboring grocery stores online and have them picked up and delivered to their home by an Instacart personal shopper. The issue statement for the Instacart competition is to forecast which previously purchased products will be in a user's next order based on their purchase history, which is a complete temporal-based data of each customer. Customers prior purchase a product and reorder as a label has been given in the dataset. We must solve the issue statement by studying the customers prior ordering pattern. According to the data, there is no cold start problem where we have to anticipate the product for new consumers because all of the user information is already in the data. The goal was to anticipate supermarket reorders based on a user's purchasing history (a collection of orders and the products purchased inside each order): which of their previously purchased products would they repurchase in their next order? The issue is distinct from the general recommendation problem, in which we frequently encounter the challenge of making predictions for new people and goods that we haven't seen before. A movie site, for example, may need to suggest new films and provide recommendations for new users. The problem's sequential and time-based aspect adds to its intrigue: how can we account for the time since a user last purchased an item? Do customers have distinct buying habits, and do they buy various goods at different times of the day? The F1 assessment criteria used by the competition ensures that our models have good precision and recall.

**Do you think there were other models they should have used? Justify your answer.**

Initially when loaded all the features to XGBoost my threshold based F1Score was 0.28 which was very low. So, to improve it I decided to get the feature importance which can help me to reduce the dimensionality of the data points. To compute this, I performed cross-validation using the LightGBM and XGBoost models. I got better results with LightGBM. LightGBM performs fast and better so I have done random search hyperparameter to get optimal score. Rather than spending more time on parameter-tuning XGBoost, I moved to LightGBM, which I’ve found to be much faster. I again opted for the random forest approach with feature\_fraction=0.8. I performed a similar grid search to the XGB approach, changing learning\_rate, num\_leaves of each tree (comparable to max\_depth for XGBoost, since LightGBM grows trees leaf-wise), and n\_estimators for the overall forest, though the best results were found with learning\_rate=0.1.

**Reference:**

[1] Kaggle Team (2020, Jan 07). Instacart Market Basket Analysis. *Kaggle Blog*. https://medium.com/kaggle-blog/instacart-market-basket-analysis-feda2700cded

**Module 3 Discussion:**

In the Tesla video (https://www.youtube.com/watch?v=YBvcKtLKNAw&ab\_channel=LexFridman), they are using multiple models to attempt to improve their self-driving features in their vehicles. One of those models is a neural network model. From the readings and from what you have learned about models, what are your thoughts/concerns of vehicles that are driving based on model predictions? Make an initial post connecting this week’s learning to the prompt.

Tesla's Neural Network will be able to interpret and identify images better due to the new way the neural network will be processing individual pieces of data. Machine learning engineers often run image recognition tests to see if their AI can differentiate between objects and organisms in the real world. Deep neural networks analyze on-car camera feeds for roads, signs, cars, obstacles, and people. Tesla's self-driving team needed a very efficient and well-designed neural network to make the most out of their high-quality dataset. The company created a hierarchical deep learning architecture composed of different neural networks that process information and feed their output to the next set of networks. It will be interesting to how the technology fares against the test of time. The capacity to predict the motions and behaviors of cars, pedestrians, and cyclists a few seconds ahead of time is known as a prediction. One of Waymo's top developers for years, recently claimed that "none has achieved" full autonomy "because today's software is not good enough to foresee the future."

The most common cause of autonomous vehicle failure is wrongly forecasting the behavior of neighboring cars and pedestrians. Tesla's fleet of roughly 500,000 automobiles is an excellent resource in this regard. When a Tesla makes an inaccurate prediction regarding a car or a pedestrian, it can save a data snapshot to upload and add to Tesla's training set later.  Rather than uploading a video, Tesla may be able to upload an abstract representation of the scene created by its computer vision neural networks (where objects are depicted as color-coded cuboid shapes and pixel-level information is discarded). The bandwidth and storage needs for transferring this data would be drastically reduced as a result. Unlike images used to train object detection, a prediction neural network can learn correlations between the past and future just by looking at temporal sequences of events. Any recording has information about what behavior comes before what behavior.

**Reference:**

[1] Lex Fridman. (Sep 25, 2018). Arguing Machines: Tesla Autopilot vs Neural Network. YouTube. Retrieved from https://www.youtube.com/watch?v=YBvcKtLKNAw&ab\_channel=LexFridman