DEGREE PROJECT IN TECHNOLOGY,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2016

American Football

A Markovian Approach

JOAKIM LARSSON

HENRIK SJÖKVIST

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF ENGINEERING SCIENCES

American Football

A Markovian Approach

J O A K I M L A R S S O N
H E N R I K S J Ö K V I S T

Degree Project in Applied Mathematics and Industrial Economics (15 credits)
Degree Progr. in Industrial Engineering and Management (300 credits)
Royal Institute of Technology year 2016
Supervisors at KTH: Henrik Hult, Jonatan Freilich
Examiner: Henrik Hult

TRITA-MAT-K 2016:23
ISRN-KTH/MAT/K--16/23--SE

Royal Institute of Technology
SCI School of Engineering Sciences

KTH SCI
SE-100 44 Stockholm, Sweden

URL: www.kth.se/sci

Abstract

This bachelor's thesis in applied mathematics & industrial economics is an attempt to model drives in American football using Markov chains. The transition matrix is obtained through logit regression analysis on historical data from the NFL. Different outcomes of drives are modelled as separate absorbing states in the Markov chain. Absorption probabilities are calculated representing the probabilities of each outcome. Results are tested against a Markov chain with the transition matrix based on frequency analysis. Three scoring rules unanimously declare the regression based model to be superior.

The application of the model pertains to live sports betting. With the insight provided by the Markovian model, a bettor should be able to make statistically informed betting decisions. The prospect of creating a start-up based on the Markovian betting model is discussed.

Keywords: *Markov Theory, Probabilistic Forecasting, Logit Regression, American Football, Sports Betting, Sports Science*

Abstract

Denna kandidatuppsats i tillämpad matematik & industriell ekonomi är ett försök till att modellera *drives* i amerikansk fotboll med hjälp av Markovkedjor. Övergångsmatrisen fås genom logit-regressionsanalys av historisk data från NFL. Olika utfall av drives modelleras som separata absorberande tillstånd i Markovkedjan. Absorptionssannolikheter beräknas, vilka representerar sannolikheterna för de olika utfallen. Resultaten testas mot en Markovkedja där övergångsmatrisen fås genom frekvensanalys. Tre olika poängregler föredrar enhälligt den regressionsbaserade modellen.

Modellens tillämpning berör sportbetting. Med hjälp av Markovmodellen bör en spelare kunna ta statistiskt underbyggda beslut i deras betting. Möjligheterna att skapa ett företag baserat på Markovmodellen diskuteras.

Nyckelord: *Markovteori, Probabilistisk Prognostisering, Logit-regression, Amerikansk Fotboll, Sportbetting, Sportvetenskap*

Contents

1 Project Description

1.1 Markovian Model of American Football

The aim of this thesis is to create a Markovian model capable of predicting the outcome of *drives* in American football. The focus lies on games in the National Football League (NFL). It is hypothesized that European sports-betting companies are underinformed with regards to American football and thus are inaccurate in their listed odds. It is believed that a Markovian model can be used to achieve higher accuracy in predicting outcomes of American football games as it pertains to live betting. The ambition is to create a model that is usable in real-time for a person watching an American football game live on television. With the statistical insight provided by the model, the person should be able to make statistically informed decisions in their live betting. The model could also be used by betting companies to improve the accuracy of their odds. Probability theory has its historical roots in gambling, and indeed, even today this is the application of probability theory in this thesis.

The game of American football lends itself nicely to discrete mathematical modelling as it is highly sequential. An introduction to the basics of American football will be included in the Theoretical Background section of the thesis.

1.1.1 Problem Formulation & Research Questions

The mathematical part of this thesis will attempt to adress two reseach questions believed to carry great importance in developing a model which can be used for live betting.

- *Can a regression based Markovian model achieve a higher prediction accuracy compared to a model based on frequency analysis when applied to an American football game?*
- *Which factors impact transition probabilities and how should they be integrated into the model?*

1.2 Industrial Management Application of Model

The prospect of creating a monetizable business based on the mathematical model is also discussed. The product is envisioned as an online application where the mathematical model is used to tell the user whether live-odds given by a betting company are favorable or not. The premise is that the model supersedes the models of the bookmakers and takes advantage of presumptive flaws thus enabling the user to bet when odds are in the user's favor. The possibilities and challenges of starting a technology-based company are explored. A strong focus lies on entrepreneurship, topics such as financing, marketing & monetization are discussed.

2 Literature Review

Previous work in Markovian modelling has been conducted on a wide array of different sports. Basketball in particular, seems to be a popular subject for Markovian modelling.[10, 13, 33, 38] Other examples of sports that have been modelled are soccer,[30, 8] baseball[3] and, as in this thesis, American football. American football is particularly well-suited for Markovian modelling due to the intrinsic sequentiality and memoryless properties of the game. These are critical characteristics for successful Markovian modelling and forecasting.

American football has been previously modelled using Markov theory, most notably by Goldner.[17] In unison with this thesis, Goldner models individual drives as Markov chains and employs Markov theory to determine probabilities of different outcomes. The approach of this thesis is similar to that of Goldner but differs in that Goldner models each play as a transition and uses frequency analysis to obtain the transition matrix. In this thesis, transitions are only made when a first down is earned or when the process is absorbed. Furthermore, the model presented in this thesis uses regression analysis instead of frequency analysis to find the transition matrix. The reasoning behind this is explained further in later sections of the thesis but essentially the changes made for this thesis allows for a smaller yet more dynamic transition matrix. Goldner's model does not allow for any dynamic input parameters; the probabilities generated by Goldner's model are the same regardless of which teams are playing, what the weather conditions are, which team has home field advantage, etc. Goldner uses a 349×349 transition matrix for the Markov chain. The model presented in this thesis *is* able to adapt to specifications such as those mentioned above, whilst only requiring a 12×12 transition matrix. The drawback of the model presented here is that it is only able to make predictions whenever a first down is earned.

3 Theoretical Background

3.1 American Football

In this section a short introduction to the basic rules of American football is given in order to familiarize the reader with the concepts on which the model is based.[1]

American football is a team sport, in which two teams of eleven players each attempt to score more points than the opposing team. The game is played with an ovoid-shaped *football*. The team in control of the football at any given time is known as the *offense*. The offense attempts to move the football down the field by passing the ball or running with it. When the forward movement of the ball is stopped, that play is declared dead. The offense will then regroup and execute a new play from the spot where the ball stopped. This creates

[1] American football strategy or theory is not discussed however. Though interesting, these topics lie well beyond the scope of relevancy for this thesis.

the sequentiality which makes Markovian modelling appropriate. The opposing team, the *defense,* attempts to prevent the offense from moving the ball and scoring points.

A game of American football lasts for 60 minutes of playing time split into four *quarters.* The football *field* is 120 yards long and 53.33 yards wide. At each end of the field is a ten yard long area known as the *end zone.* Yard markers on the field indicate the distance from the nearest end zone. At the back of each end zone is a tall fork-shaped *goalpost.* [18]

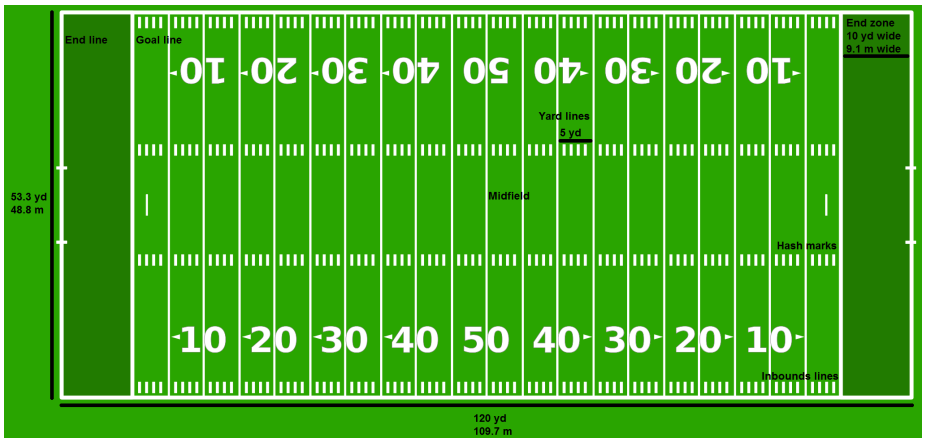

Figure 1: An American football field

Points are scored in four possible ways.

- A *touchdown* (TD) is scored when a ball is caught in, or advanced into the opposing team's end zone. A touchdown is worth six points.
- A *field goal* (FG) is scored when the ball is kicked through the uprights of the opposing team's goalpost. A field goal is worth three points.
- After scoring a touchdown, the scoring team attempts a *point-after-touchdown* (PAT). The team is given the option of either kicking a field goal worth one point, or playing a single regular play to score a touchdown worth two points.
- A *safety* is scored when a player carrying the ball is tackled within their own end zone. A safety is worth two points, which are awarded to the defense.

During play, the offense is given a sequence of four attempts to move the ball a total of ten yards forward. These attempts are known as *downs.* If the offense succeeds in moving the ball ten or more yards in four or less downs they are awarded a new set of four downs and the objective of moving the ball another ten yards. Then the next play is known as a *first down.* If the offense fails in

moving the ball ten yards the ball is *turned over* to the opposing team. This is known as a *turnover on downs.* In most cases if an offense has exhausted three of its downs, they will want to avoid the risk of turning the ball over by failing to convert the fourth down. Instead, they may choose to kick a field goal or *punt* the ball. A punt is when a player on the offense kicks the ball deep down the field. The defense then recovers the punt and attempts to run the ball back as far as possible.

If the team on defense obtains possession of the ball during play they become the offense. The team on offense which lost control of the ball is then forced to play defense. A number of different scenarios cause the teams to switch possession of the ball. If a thrown ball is intercepted, or if a fumbled ball is recovered by the defense, we refer to it as a *turnover.* If the offense scores points, or if the offense punts the ball, then possession of the ball is also switched.

A *sequence* is defined to be all the plays from a first down until the team either earns a new first down, punts the ball, turns the ball over or scores points. A *drive* is defined as all the sequences from a first down until the team either punts the ball, turns the ball over or scores points. Again, the objective of this thesis is to determine the probabilities for the outcomes of a drive.

3.2 Probability Theory

3.2.1 Probability Spaces and Random Variables

Definition 3.2.1.1 A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where[29]:

(i) Ω is the set of all possible outcomes of a random event, it is known as the *sample space.*

(ii) $\mathcal{F}$ is a collection of subsets of Ω structured as a σ-algebra (or σ-field):

1. $\emptyset \in \mathcal{F}$
2. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$
3. $A_1, A_2, \ldots \in \mathcal{F} \Longrightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

(iii) $\mathbb{P}$ is a probability measure, i.e. a function which associates a number $\mathbb{P}(A)$ to each set $A \in \mathcal{F}$ such that:

1. $0 \leq \mathbb{P}(A) \leq 1$
2. $\mathbb{P}(\Omega) = 1$
3. $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ for any sequence $A_1, A_2, \ldots$ of pairwise disjoint sets in $\mathcal{F}$

Definition 3.2.1.2 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *random variable* is a function $X : \Omega \to \mathbb{R}$ such that for every set $A \in \mathcal{B}$,

$$X^{-1}(A) = \{\omega : X(\omega) \in A\} \in \mathcal{F}$$

where $\mathcal{B}$ is the Borel σ-algebra over $\mathbb{R}$.[22]

3.2.2 Stochastic Processes

Definition 3.2.2.1 A *stochastic process* is a family of random variables $X(t)$ defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$[29],

$$\mathbf{X} = \{X(t)|t \in T\}.$$

The set T is known as the *parameter set.* $T \subseteq \mathbb{N}$ in the case of a *stochastic process in discrete time* and $T \subseteq \mathbb{R}$ in the case of a *stochastic process in continuous time.* The set $\mathbb{S}$ of possible values that $X(t)$ may take is known as the *state space.*

3.3 Markov Theory

3.3.1 Fundamental Definitions

Definition 3.3.1.1 A *Markov chain* is a stochastic process in discrete time $\mathbf{X} = \{X_n|n \in \mathbb{N}_0\}$ such that

$$\mathbb{P}(X_{n+1} = s_{n+1}|X_0 = s_0, X_1 = s_1, \ldots, X_n = s_n) = \mathbb{P}(X_{n+1} = s_{n+1}|X_n = s_n)$$

for all $n \in \mathbb{N}_0$ and all *states* $s_0, s_1, \ldots, s_n, s_{n+1} \in \mathbb{S}$ where $\mathbb{S}$ is the state space of $\mathbf{X}$.[42]

The *transition probability from i to j,* p_{ij} is defined as

$$p_{ij} = \mathbb{P}(X_n = s_i|X_{n-1} = s_j),\ s_i, s_j \in \mathbb{S}.$$

The *transition matrix* $\mathbf{P}$ is the matrix $(p_{i,j})_{s_I, s_j \in \mathbb{S}}$ consisting of the transition probabilities between corresponding rows and columns

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{22} & p_{22} & p_{23} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

3.3.2 Absorption

A state s_i of a Markov chain is called *absorbing* if $p_{ii} = 1$, i.e. if it is impossible to leave the state once entered. A Markov chain is called *absorbing* if one or more of its states are absorbing. In an absorbing Markov chain, a state which is not absorbing is called *transient.*[19] Thus the state space $\mathbb{S}$ can be split into a set $\mathbb{A}$ of absorbing states and a set $\mathbb{T}$ of transient states, such that $\mathbb{S} = \mathbb{A} \cup \mathbb{T}$.[12]

An absorbing Markov chain with a finite state space will absorb the process with probability 1 in finitely many steps. Let $a_{i,j}$ denote the *absorption probability,* the probability of absorbtion in state j given that the process currently is in state i, $s_j \in \mathbb{A}, s_i \in \mathbb{T}$. The absorption probabilities can be found by solving the following equation system for all $s_j \in \mathbb{A}, s_i \in \mathbb{T}$

$$a_{i,j} = p_{i,j} + \sum_{k \in \mathbb{T}} p_{i,k} a_{k,j}$$

The equation system can be expressed on matrix form. Let $\mathbf{Q} = (p_{i,j})_{s_i \in \mathbb{T}, s_j \in \mathbb{A}}$ and $\mathbf{R} = (p_{i,j})_{s_i, s_j \in \mathbb{T}}$. Furthermore, let $\mathbf{A} = (a_{i,j})_{s_i \in \mathbb{T}, s_j \in \mathbb{A}}$. Then

$$\mathbf{A} = (\mathbf{I} - \mathbf{R})^{-1}\mathbf{Q} \tag{1}$$

where $\mathbf{I}$ is the identity matrix.[12]

3.4 Regression Analysis

Regression analysis is a mathematical tool used to examine statistical data and investigate if relationships exist between variables.[37] The fundamental mathematical model indicates a particular relationship between a *response variable* and *explanatory variables* (or *covariates*).[23] The specified relationship depends on underlying assumptions regarding how the variables are related.

3.4.1 Fundamental Definitions

A *response variable* y is a single value from a defined category (e.g. *age*), generated either by observation or by an experiment. The value of the response variable is explained through a mathematical relationship by *covariates* x, to an extent. Covariates are explicitly valued from single data categories and can be categorized as *observational* or *experimental*. The relationship between the response variable and the covariates is generally not perfectly modelled, hence an error term ε is included. The error term is known as the *residual*.[23]

A *regression* implies the use of a mathematical method to fit the data yielded by the covariates to the data of the response variable such that the residual, or error, is reduced. In general, the regression generates estimated values of coefficients linked to each covariate. The estimation elucidates the impact of each covariate upon the value of the respone variable. In *linear regression,* the response variable is modelled by a linear combination of the covariates according to:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

where $\beta_i \in \mathbb{R}, i = 0, 1, \ldots, n$ are the coefficients to be estimated. β_0 is known as the *intercept*.

3.4.2 p-value

The *p-value* is defined as the probability $\mathbb{P}(\mathbf{Z} > z)$ where z is an outcome of the random variable $\mathbf{Z}$. A high p-value indicates that an estimate or a set of estimates could be equal to zero and thus does not carry enough significance to be included in the model.[23]

3.4.3 Logit

The Logit model is a suitable regression model for estimates of probability. It estimates the probability of the occurence of a singular event. The model is

defined as follows:[25]

$$y_i = \frac{e^{\mathbf{x_i}\beta}}{1 + e^{\mathbf{x_i}\beta}} = p(\mathbf{x_i}, \beta) \tag{2}$$

where y_i is an observation of the outcome of an event A. y_i is a binary variable assuming values:

$$y_i = \begin{cases} 1, & \text{if } A \text{ occured} \\ 0, & \text{if } A \text{ did not occur} \end{cases}$$

Furthermore, n denotes the number of observations, $\mathbf{x_i}$ is the vector with values of the covariates as pertained to the sample i. β is the vector with coefficients corresponding to each covariate. The regression is performed by maximizing the *log-likelihood* function

$$\ln(L) = \sum_{j=1}^{n} ln[(2y_i - 1)p(\mathbf{x_i}\hat{\beta}) + 1 - y_i]$$

with regard to $\hat{\beta}$.

3.4.4 Model Testing

Covariates in a regression model are chosen ad hoc by the modeler with regard to contextual factors. Hence, regression models and estimates are generally tested to ascertain whether or not covariates are significant. A common test is to try the *null hypothesis*, i.e. to test if the estimated coefficient is statistically significantly not equal to zero. A null hypothesis test can be performed in several ways.

The Log-Likelihood Ratio Test. Let $\ln(L_*)$ be the log-likelihood function with the coefficients of covariates to be tested set to zero. Let r be the number of *restrictions,* i.e. the number of coefficients set to zero. Then,

$$\mathcal{L} = 2\ln(L) - 2\ln(L_*)$$

is approximately an outcome of a $\chi^2(r)$-distributed variable. The p-value is computed: $p = \mathbb{P}(\chi^2(r) > \mathcal{L})$. If $p > \alpha$ where α is some specified tolerance level then the full model is rejected in favor of the restricted model.

The Wald Test. Let $\hat{\beta}_i$ be an estimated coefficient for a covariate x_i. Then,

$$\mathcal{W} = \frac{\hat{\beta}_i^2}{\mathrm{Var}(\hat{\beta}_i)}$$

is approximately an outcome of a $\chi^2(1)$-distributed variable under the null hypothesis.[2] The outcome is used to determine a p-value: $p = \mathbb{P}(\chi^2(1) > \mathcal{W})$. If $p > \alpha$ where α is some specified tolerance level then the covariate is excluded from the model.[4]

[2]Note that the Wald test can be used to test the null hypothesis for a set of several coefficient estimates. Such a test will not be performed in this thesis, however.

AIC. The *Akaike Information Criterion* (AIC) is an estimate of the information loss which occurs when using a regression model to approximate the relation between explanatory and response variables. Let n be the number of observations, k the number of covariates in the model tested, and $|\hat{\varepsilon}|^2$ the square of the Euclidean norm of the residual. Then $n\ln(|\hat{\varepsilon}|^2)+2k$ is an estimate of the information loss of a linear model.[23] In the particular case of a logit regression the AIC is:

$$-2\ln(L)+2k$$

where $\ln(L)$ is the log-likelihood functon. The objective is to find the model which minimizes the AIC-value.

3.5 Decision Theory

3.5.1 Scoring Rules

Statistical analysis can be utilized in order to produce probabilistic forecasts for future events. *Scoring rules* assign a numerical score to the forecast based on the predictive probability distribution P and on the realized event ω. The scoring rule, or *scoring function* is a function $S(P,\cdot)$ taking values in the extended real line.

Definition 3.5.1.1 Consider a probability space $(\Omega,\mathcal{F},P)$, and let $\mathcal{P}$ be a convex class of probability measures on $(\Omega,\mathcal{F})$ such that $P\in\mathcal{P}$. A scoring rule is any function $S:\mathcal{P}\times\Omega\to\bar{\mathbb{R}}$ such that $S(P,\cdot)$ is measurable with respect to $\mathcal{F}$. If the forecast is P and ω materializes, the score is $S(P,\omega)$. [16]

The scoring rules used in this thesis are *positively oriented,* meaning that the optimal score is the maximum. Thus a prediction which scores higher than another prediction is deemed more successful.[3]

Definition 3.5.1.2 Let $P,Q\in\mathcal{P}$. A scoring rule is said to be *strictly proper* if the expected score under Q when the forecast is P is uniquely maximized by $P=Q$.

Usage of proper scoring rules encourages the forecaster to always quote their true belief in the forecast. In this thesis, three different strictly proper scoring rules will be used:

- The *logarithmic scoring rule.* $L(P,\omega)=\ln(P(\omega))$.[4]
- The *quadratic scoring rule.* $Q(P,\omega)=2P(\omega)-\sum_{\omega\in\Omega}P(\omega)^2$
- The *spherical scoring rule.* $S(P,\omega)=\frac{P(\omega)}{\sqrt{\sum_{\omega\in\Omega}P(\omega)^2}}$

[3]There exist scoring rules which are negatively oriented, one such example is the *Brier score.* The Brier score is closely related to the Quadratic scoring rule, used in this thesis.

[4]The logarithmic scoring rule is closely related to the entropy H of a discrete probability distribution. $H(X)=-\sum_{i=1}^{n}P(x_i)\ln P(x_i)$.

3.6 Sports Betting

In sports betting an individual, the *bettor,* wagers an amount of money, the *stake*, on the outcome of a sports event. The bets are placed with a *bookmaker.* The bookmaker is an individual or organization that facilitates the betting.

3.6.1 Odds

If the bettor has placed a bet and is correct in predicting the event, then the bettor earns the stake back plus a profit for being correct. The profit depends on a metric known as *odds.* The odds are predetermined prior to the placement of the bet. In this thesis the odds will be expressed as decimal values greater than 1, known as *decimal odds.* Decimal odds are the most common form of expressing odds in continental Europe.[5]

Let $\kappa > 1$ be the decimal odds for a bet. Furthermore let $s > 0$ be the stake of the bet. Then the (gross) returns r for the bettor can be expressed as:

$$r = \begin{cases} \kappa s, & \text{if the bet is correct} \\ 0 & \text{if the bet is incorrect} \end{cases} \tag{3}$$

The sports betting odds are not to be confused with the concept of *statistical odds.* In statistics, the odds in favor of an event is the ratio of the probability that the event will happen to the probability that the event will not happen. Clearly this is not equivalent to any of the definitions of sports betting odds given in this thesis. In this thesis, *odds* will refer to sports betting decimal odds.

4 Methodology

Play-by-play data from the NFL is gathered and a dataset is procured. The dataset is used in a logit regression model to estimate transition probabilities for a Markov chain. Absorption probabilities are calculated from the Markov chain and displayed in a user interface relaying relevant betting information.

4.1 Markovian Modelling

Each drive is modelled as a Markov chain. In the Markovian model a transition is made each time a new first down is earned, or the drive ends. This demands

[5] In the United Kingdom and Ireland, *fractional odds* are the most common. Fractional odds κ_f are expressed as rational numbers that exhibit *net* returns. The gross returns are given by $r = \begin{cases} (\kappa_f + 1)s, & \text{if the bet is correct} \\ 0 & \text{if the bet is incorrect} \end{cases}$.

In America, *moneyline odds* are favored. Moneyline odds κ_m are quoted as either positive or negative integers. If the odds are positive, they are quoting the net returns on a wager with $s = 100$. If the odds are negative, they are quoting the stake size required in order to gain a net return of 100.

an explanation. One might consider the natural way of modelling a drive to be to have every play be a transition. However as teams play dramatically differently on first downs compared to fourth downs, this leads to complications. One way to remedy this might be to multiply the number of states by four, having one set of states for each possible current down. This not only leads to a very large transition matrix, but more importantly dilutes the statistical data far too much. Another suggestion might be to instead have four different transition matrices, one for each down. However, this causes the Markov chain to be *time-inhomogeneous*. The analytical solutions for absorption probabilities presented in **3.3.2** require the Markov chain to be *time-homogeneous*, i.e. have a constant transition matrix. If time-homogeneity is abandoned then numerical methods must be resorted to in order to find the absorption probabilities.

Instead, the process transitions with each new series. This significantly simplifies computations. Meanwhile the drawbacks are quite limited. The only individual plays that don't affect the process are those that have small or no impact on the ball position. Any big plays will result in a new sequence, and thus be modelled by the Markov chain. The other drawback is that the person betting will only have statistical data for first down scenarios. The model will not be able to explain changes in probabilities of outcome from a first to a second down.

4.1.1 Definition of States

The transient states consist of sections of the football field in which a first down is obtained. The field is divided into eight sections, each measured in yards from opponent goal line. They are denoted by index 1-8 in the following order:

1. 100-90 yards. Here the team is closely backed up against their own end zone. There is more pressure and less space for the offense to work with which should increase the probability of turnovers and punts.

2. 90-60 yards. The objective in this area is simply to advance the ball forward. Touchdowns are very rare from this far away and field goals are impossible.

3. 60-40 yards. The ball is at midfield, very long field goals are possible but rare.

4. 40-20 yards. The team is within range for field goals which eliminates the need for punting.

5. 20-15 yards. The 20 yards closest to the opponent's end zone is known as the *red zone*. Here the probability for touchdowns is the highest, and gets increasingly higher as the distance to the end zone decreases. The red zone is divided into four sections to better represent the increasing probability of scoring as the ball gets closer to the end zone.

6. 15-10 yards.

7. 10-5 yards.

8. 5-0 yards.

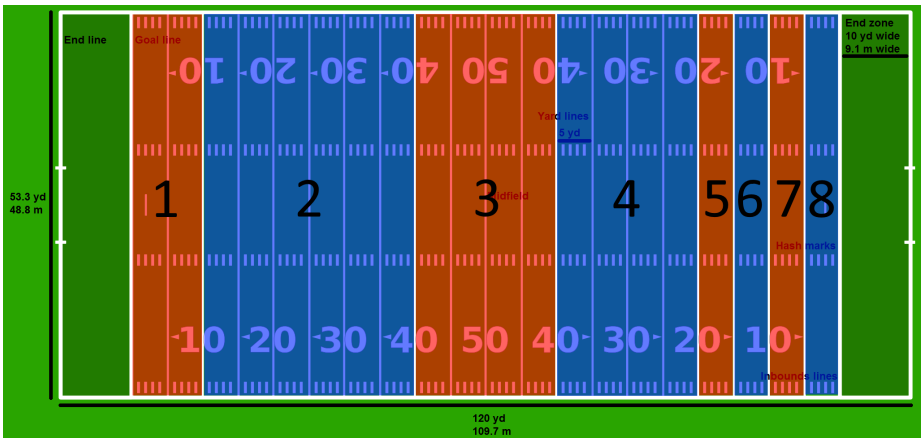

Figure 2: The transient states of the Markov chain

The possible outcomes of a drive are represented in the Markovian model as absorbing states. Thus the outcomes *touchdown, punt, field goal* and *turnover* are represented with states with indices 9, 10, 11 and 12 respectively.

4.1.2 Transition Matrix

There are twelve states in the Markovian model. The transition probabilities for each pair of states is required to form the 12×12 transition matrix. For instance, $p_{2,5}$ denotes the probability of obtaining the next first down in the area 20-15 yards from goal line given that the previous first down was obtained 90-60 yards from the goal line. $p_{6,9}$ denotes the probability of scoring a touchdown in the current set of downs given that the previous first down was obtained 15-10 yards from the goal line. Note that the process may transition to the same state twice or more times consecutively. This represents the team earning its next first down in the same section of the field as the previous first down.

4.1.3 Absorption Probabilities

The Markovian model includes four absorbing states. Each drive is guaranteed to end with an outcome represented by one of the absorbing states. Given that the ball is in a state $s_i, i \in \{1, 2, \ldots, 8\}$ it is of interest to determine the probability distribution for the four possible absorption probabilities $a_{i,9}, a_{i,10}, a_{i,11}, a_{i,12}$. [6] The absorption probabilities are obtained by solving the

[6] Note the difference between, for instance, $p_{1,9}$ and $a_{1,9}$. $p_{1,9}$ is the probability that a *sequence* which starts 100-90 yards from the goal line ends with a touchdown. $a_{1,9}$ is the

following system:

$$\mathbf{A} = (\mathbf{I} - \mathbf{R})^{-1}\mathbf{Q}$$

where

$$\mathbf{A} = \begin{bmatrix} a_{1,9} & a_{1,10} & a_{1,11} & a_{1,12} \\ \vdots & \vdots & \vdots & \vdots \\ a_{8,9} & a_{8,10} & a_{8,11} & a_{8,12} \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} p_{1,1} & p_{1,2} & p_{1,3} & p_{1,4} & p_{1,5} & p_{1,6} & p_{1,7} & p_{1,8} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{8,1} & p_{8,2} & p_{8,3} & p_{8,4} & p_{8,5} & p_{8,6} & p_{8,7} & p_{8,8} \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} p_{1,9} & p_{1,10} & p_{1,11} & p_{1,12} \\ \vdots & \vdots & \vdots & \vdots \\ p_{8,9} & p_{8,10} & p_{8,11} & p_{8,12} \end{bmatrix}$$

and $\mathbf{I}$ is an 8×8 identity matrix.

4.2 Data collection

In order to find the transition matrix for the Markov chain, a regression is performed on a dataset. This section covers how the data was collected and refined.

4.2.1 Selection of Data

The fundamental data points used in this thesis are individual downs (plays) executed in the National Football League. Games from the 2009 through 2015 *regular seasons* are used to gather play-by-play data. A regular season in the NFL consists of 256 games, played by 32 teams. Each team plays 16 games in one regular season.[7] In one game an average of 65 plays are executed by either team's offense (130 offensive plays in total).[34] Thus, roughly 33,280 plays are executed each regular season. Over the seven seasons used this equates to 232,960 plays.

For the purpose of creating and testing the model, only plays from one team are used - the Pittsburgh Steelers. The Steelers have been a consistent performer with regards to win-loss record over the seasons analyzed. Furthermore, they have had the same *Head Coach* and *Quarterback* over the seasons. The reasoning behind the team selection is that a consistent performing team might mitigate effects from complex factors such as team strength, injuries to

probability that a *drive* which starts 100-90 yards from the goal line ends with a touchdown. The former requires the offense to move the ball at least 90 yards in a single play in order not to earn a new set of downs before the touchdown. The latter allows the team to earn any number of first downs, so long as the drive eventually ends with a touchdown. Thus one expects $a_{1,9}$ to be significantly larger than $p_{1,9}$.

[7] Pre-season games and playoff games are not included in the regular season.

key players, and opponent team strength. These factors are hard to quantify and use in an estimate.

Only offensive plays by the PITTSBURGH STEELERS are included in the dataset. Defensive plays and kick-offs are assumed to be independent of offensive play performance and are discarded.

4.2.2 API & NFL.com Database

Data of all the plays is gathered through an open source API named *nflgame.[15]* The API retrieves and enables reading of NFL Game Center JSON data, a database found on www.nfl.com. The website nfl.com is the official website of the NFL and tracks all plays in real time, archiving them upon completion of a game.

Through Python code, *nflgame* retrieves all offensive plays made by the STEELERS from 2009 to 2015. These plays are categorized into sequences, i.e. all the plays from a first down until the team either earns a new first down, punts the ball, turns the ball over or scores points. Each sequence is encoded into a *comma separated values* file with information about sequence id, year, time, home field, starting yard line of the sequence, ending yard line of the sequence and binary values for whether a touchdown, punt, field goal or a turnover occured.

4.2.3 Weather Data

Weather data for all STEELERS games is obtained from www.nflsavant.com [43], an advanced statistics site for NFL games. Data is transcribed from datasets obtained from NFLsavant to the comma separated values file containing all offensive plays. The weather data used is the temperature and wind speed for each game. Temperature is measured in degrees Fahrenheit. Wind speed is measured in miles per hour. Temperature and wind speed are assumed to be constant throughout the game.

4.2.4 Defensive Data

An offense should struggle more when facing a good defense, ceteris paribus. To account for this, a defensive ranking is compiled to be used in the regression. Opposing teams to the PITTSBURGH STEELERS are ranked defensively on a year-to-year basis. The ranking is implemented by collecting all of the team's yards allowed per game and points allowed per game in a given year, and are ranked relative to all other NFL teams. The defensive ranking is then the mean of the two rankings given in Table 1. The data is collected from nfl.com. In the regression model these values are multiplied by 100. Thus the defensive ranking is a number in the interval $[0, 100]$.

Team	Pts/G	Yds/G	YardIndex	Pts/G	Combined
New York Jets	14,8	252,3	1	1	1
Green Bay Packers	18,6	284,4	0,8871308	0,79569892	0,84141486
Baltimore Ravens	16,3	300,5	0,83960067	0,90797546	0,87378806
Cincinnati Bengals	18,2	301,4	0,83709356	0,81318681	0,82514019
Pittsburgh Steelers	20,2	305,3	0,82640026	0,73267327	0,77953676
Minnesota Vikings	19,5	305,5	0,82585925	0,75897436	0,7924168
Denver Broncos	20,2	315	0,80095238	0,73267327	0,76681282
Carolina Panthers	19,2	315,8	0,79892337	0,77083333	0,78487835
Dallas Cowboys	15,6	315,9	0,79867047	0,94871795	0,87369421
Washington Redskins	21	319,7	0,78917735	0,7047619	0,74696963
New England Patriots	17,8	320,2	0,78794503	0,83146067	0,80970285
Philadelphia Eagles	21,1	321,1	0,78573653	0,7014218	0,74357917
Houston Texans	20,8	324,9	0,77654663	0,71153846	0,74404255
New York Giants	26,7	324,9	0,77654663	0,55430712	0,66542687
San Francisco 49ers	17,6	326,4	0,77297794	0,84090909	0,80694352
San Diego Chargers	20	327	0,77155963	0,74	0,75577982
Chicago Bears	23,4	337,8	0,74689165	0,63247863	0,68968514
Indianapolis Colts	19,2	339,2	0,74380896	0,77083333	0,75732115
Buffalo Bills	20,4	340,6	0,74075161	0,7254902	0,73312091
Arizona Cardinals	20,3	346,4	0,72834873	0,72906404	0,72870638
Atlanta Falcons	20,3	348,9	0,72312984	0,72906404	0,72609694
Miami Dolphins	24,4	349,3	0,72230175	0,60655738	0,66442956
Jacksonville Jaguars	23,8	352,3	0,71615101	0,62184874	0,66899987
Seattle Seahawks	24,4	356,4	0,70791246	0,60655738	0,65723492
New Orleans Saints	21,3	357,8	0,70514254	0,69483568	0,69998911
Oakland Raiders	23,7	361,9	0,69715391	0,62447257	0,66081324
Tampa Bay Buccaneers	25	365,6	0,69009847	0,592	0,64104923
Tennessee Titans	25,1	365,6	0,69009847	0,58964143	0,63986995
St, Louis Rams	27,2	372,8	0,67677039	0,54411765	0,61044402
Kansas City Chiefs	26,5	388,2	0,64992272	0,55849057	0,60420664
Cleveland Browns	23,4	389,3	0,64808631	0,63247863	0,64028247
Detroit Lions	30,9	392,1	0,6434583	0,4789644	0,56121135

Table 1: Snapshot of 2009 Defensive Rankings

4.2.5 Data Weighting by Pseudo-Samples

Careful consideration must be placed on the predictive power of historic data. Historically, the rules and the way the game is played has not changed dramatically from 2009 to 2015. However on a team level there is always a lot of change. This is mainly due to the continuous roster turnover. Pittsburgh Steelers retained only 14 out of 53 (26.4%) players from the 2011 season to the 2014 season. The highest retainment percentage belongs to the Green Bay Packers with 47.1%. It is realistic to think that the roster turnover affects team perfor-

mance on a yearly basis. Another important factor is changes to the coaching staff. Replacing the coaching staff can dramatically change the way the team plays. As mentioned though, this is not applicable to the STEELERS during the 2009-2015 time frame as they have had the same Head Coach, and changes at other coaching positions have been rare. Nevertheless, it is reasonable to believe that due to roster turnover the data from 2009 carries less predictive power than data from 2014. Thus, data is weighted to correct for this occurence.

A weight is assigned to each season according to Table 2. The observations in each year are duplicated according to the value of the weight. E.g. an observation from 2013 is counted three times as separate data points in the regression.

Season	2009	2010	2011	2012	2013	2014
Weight	1	1	1	2	3	5

Table 2: Weights associated to data from certain years.

4.3 Transition Probability Estimation

4.3.1 Frequency Analysis

The most basic approach to estimating transition probabilities is to look at the frequencies of such transitions in the dataset. A frequency analysis is performed when the regression model is unable to provide estimates, e.g. if none of the covariates are significant. A transition probability $p_{i,j}$ is estimated by counting all jumps from state s_i to s_j, and dividing by total number of jumps out of state s_i.

$$p_{i,j} = \mathbb{P}(X_{t+1} = s_j | X_t = s_i) = \frac{n_{i,j}}{\sum_k n_{i,k}}$$

4.3.2 Regression Hypotheses

The main tool for estimating the transition probabilities in this thesis is the logit regression model. The regression is performed using the software R, which is a software designed for statistical analysis. The software uses its default *glm()* function to perform a logit regression. To align the regression with the Markov matrix, the sequences from the comma separated values file are divided into 8 different data sets by the different areas of the field, e.g. all sequences which started in the region 90-60 yards from the opponent endzone are represented in a set. Thus each data set corresponds to one of the transient states in the Markov matrix.

The logit regression adds dynamism to the model. There is a belief that factors such as temperature or opponent team strength impact the transition probabilities. The regression answers the question regarding how large impact such factors have, if any at all. As stated in the theoretical framework, finding appropriate covariates is done by an analysis of the context.

Hypothesis 4.3.2.1: *Temperature has an impact on some transition probabilities.*

A majority of American football games are played outdoors. Hence, players are subjected to varying weather conditions. Colder temperatures make the football harder to grip, increasing the difficulty of passing and catching the footabll. Subsequently, the offense should run the ball more and pass less. A run generally gains less yardage than a pass and consequently a run-heavy game yields less scores and more punts. This is because the defense knows the propensity for the offense to run due to the cold weather and adjust their defense to counter run plays. Hence, it can be expected that cold weather impacts punt probabilities positively and advancement probabilities as well as touchdown probabilities negatively. Moreover, field goal probabilities should increase too as advancement is harder. Finally, turnover probability can be expected to stay constant because fumble probabilities should increase with the increased amount of runs but should be compensated by less interceptions as passing decreases. Warm weather should have the opposite effect. The coefficient to the covariate is denoted β_{Temp}.

Hypothesis 4.3.2.2: *Wind has an impact on some transition probabilities.*

Any time the ball travels through the air it can be affected by wind. Thus, passing and kicking is more difficult if the wind speed is high. Interceptions and thus the number of turnovers would probably increase and field goal probability should be reduced. This coefficient is denoted β_{Wind}.

Hypothesis 4.3.2.3: *Home Field Advantage has an impact on some transition probabilities.*

A debated topic is how being on home field affects the performance of a team. Research suggests that playing on home field has a positive impact on team performance.[11]. Hence, it can be expected that touchdown and advancement probabilities increase with home field advantage whereas punt and turnover probabilities decrease. The coefficient is denoted β_{HFA}.

Hypothesis 4.3.2.4: *The game being in the fourth quarter has an impact on some transition probabilities.*

Football could be described as a patient game. A game typically involves multiple scores (touchdowns and field goals). Teams do not change their strategy when the opponent scores but rather assume a methodical approach with low risk taking. Generally, this approach changes to a riskier one in the fourth quarter if the team is trailing in points since there is limited time left to score. Conversely, a team in the lead typically tries to be more conservative. However, that approach normally deviates less from the original approach than the risky strategy. Hence, shorter advancements should be less likely in favor of an increase in longer advancements. Likelihood of Touchdowns, Field Goals and Turnovers should increase while Punts should decrease. Q4 is a dummy covariate and the coefficient is denoted β_{Q4}.

Hypothesis 4.3.2.5: *The defensive strength of the opposing team has an impact on some transition probabilities.*

The team which is on offense will have a harder time to advance the ball and score points if the opposing team has a good defense. Thus, defensive strength should negatively impact touchdown, field goal, and advancement probabilities and increase turnover and punt probabilities. The coefficient is denoted β_{ODS}.

Regression Model

The basic logit model which is used in the regression is thus:

$$y = \frac{e^{\mathbf{x}\beta}}{1+e^{\mathbf{x}\beta}} = \frac{e^{\beta_0+x_{Temp}\beta_{Temp}+x_{Wind}\beta_{Wind}+x_{HFA}\beta_{HFA}+x_{Q4}\beta_{Q4}+x_{ODS}\beta_{ODS}}}{1+e^{\beta_0+x_{Temp}\beta_{Temp}+x_{Wind}\beta_{Wind}+x_{HFA}\beta_{HFA}+x_{Q4}\beta_{Q4}+x_{ODS}\beta_{ODS}}}$$

4.3.3 Logit Model Testing

Each transition probability in the Markov transition matrix is attempted to be estimated by the use of the basic logit model. The model is tested after each regression to establish whether covariates are significant or not. Non-significant covariates should be excluded from the model for that particular transition probability. It is of course possible that certain covariates only carry significance in sections of the Markov chain. The following algorithm is used to test and determine the appropriate model for each regression:

1. Each transition is checked to identify if an event is non-occurring. If it is non-occuring then the transition probability is set to zero.

2. The Wald-test is performed on each of the covariates to establish a p-value with respect to the $\chi^2(1)$ distributed $\mathcal{W}$-statistic. All covariates below the significance level of 90% are discarded, i.e. all covariates with a p-value higher than 0.1. This implies that there is at most a 10% risk of a false rejection of the null hypothesis, i.e. at a risk of 10% it can be claimed that the covariate is different from zero. The AIC is also calculated for the full model to be used for comparison later in the testing algorithm.

3. If all of the covariates are discarded the loglikelihood-ratio test is performed. The test is performed with a null restriction, i.e. the reduced model includes only the intercept. A p-value is calculated with respect to the $\chi^2(5)$ distributed $\mathcal{L}$-statistic. The entire regression model is discarded in favor of the frequency analysis, if the p-value is higher than 0.1. Otherwise, the model is reduced to the intercept alone and the regression is performed again to establish the beta-value.
   If covariates remain, the regression is performed again without the discarded covariates. The Wald-test is completed again to ascertain that previous significance levels of covariates were not due to misspecifications or other noise in the full model. Any covariate below the significance level of 90% is discarded and this step is repeated.

4. The AIC-value is calculated for the remaining model and compared with the full model. The reduced model is used as an estimate of the transition probability if the AIC-value is less than the full model. Otherwise, the AIC for different combinations of the remaining covariates are computed to see if a model with a lower AIC can be found and used. If such a model is not found, the initial remaining model is still used with a full understanding that it might not approximate reality as well as another model.

4.3.4 Markovian Model Testing

A test environment is set up to test the Markovian model. The environment is set up such that 2015 data is separated and not included in the logit regression. The model will attempt to predict results in the 2015 season using historical data from the 2009-2014 seasons.[8] The test environment models the situation a bettor is in at the beginning of a season, when only data from past seasons is available.

Every offensive drive by the PITTSBURGH STEELERS from the 2015 season is included in the test environment, along with data regarding the temperature, wind speed, opponent, quarter and home field advantage. For each first down, a transition matrix is estimated using the logit estimates, and an absorption probability distribution is computed. The prediction is scored using the logarithmic, quadratic and spherical scoring rules. This is iterated for every first down over the entire season and mean scores are computed for each of the scoring rules.

The process is then repeated but instead of using a transition matrix from the logit regression, a transition matrix based on frequency analysis is used. The transition matrix based on frequency analysis is the same for every drive, regardless of covariate values. The mean scores from the scoring rules are again computed and compared to those from the regression based transition matrix.

4.4 Results Interpretation

This section covers how the results of the Markovian computations are presented to the user. An analysis of the results and accuracy of the model is found in section **5.**

The absorption probabilities must be put into context in order to be of actual use for the bettor. The model can not determine whether or not a bet is a good investment without knowing the odds that are available to the bettor. One might expect the probability $a_{8,9}$ of scoring a touchdown given that a first down is earned less than 5 yards from the opponent's end zone to be very high. However the odds available to the bettor for a touchdown are likely to be very

[8] The most obvious way of testing the accuracy of the predictions made by the Markovian model would be to watch live NFL games and bet money according to the model's suggestions. If a profit is made, then the model is considered successful. However, this thesis was written during the spring of 2016, the NFL season spans September through February. As such, there were no live games available for model testing. This will be discussed further in **5.1.1**.

low for the same reason. Thus betting on a touchdown to be scored might not be a statistically sound decision. For this reason it is of interest to the bettor to know the minimum odds required for an expected profit.

Consider the previous example of betting on the outcome of a drive being a touchdown when a first down is earned less than 5 yards from the opponent's end zone. Let s be the size of the stake. Let κ be the odds available to the bettor for this bet. From **(3)** the returns on the bet are given by

$$r = \begin{cases} \kappa s, & \text{if the bet is correct} \\ 0 & \text{if the bet is incorrect} \end{cases}.$$

r is a random variable. The expected value of r is given by $\mathbb{E}[r] = \kappa s \times \mathbb{P}(\text{bet is correct}) + 0 \times \mathbb{P}(\text{bet is incorrect}) = \kappa s \times a_{8,9}$. A profit is made if $r > s$. Thus, in order for the bet to have an expected profit the following must hold:

$$\begin{aligned} \mathbb{E}[r] &> s \\ \kappa s a_{8,9} &> s \\ \kappa a_{8,9} &> 1 \\ \kappa &> \frac{1}{a_{8,9}} \end{aligned}$$

If the odds available to the bettor are greater than the multiplicative inverse of the probability of the bet being correct, then the bet carries an expected profit and should be exercised.

5 Results

5.1 Logit Regression Results

The selected model for each transition is given here, determined in the software R using the algorithm presented in **4.3.3.** Beta values for each regression is presented in Table 3

From (row)	To (column)	Intercept	Temp	Wind	HFA	Q4	OppDSth	Från (rad)	Till (kolumn	Intercept	Temp	Wind	HFA	Q4	OppDSth
100-91 (1)	100-91 (1)	-3,3557	0	0	0	0	0	20-16 (5)	100-91 (1)	non	occurrence				
100-91 (1)	90-61 (2)	-1,92642	0,0343	0	0	0	0	20-16 (5)	90-61 (2)	non	occurrence				
100-91 (1)	60-41 (3)	-5,86543	0	0,1562	0	2,50922	0	20-16 (5)	60-41 (3)	non	occurrence				
100-91 (1)	40-21 (4)	19,5609	0	0	0	0	-0,3398	20-16 (5)	40-21 (4)	-17,58344	0	0	0	0	0,18182
100-91 (1)	20-16 (5)	non	occurrence					20-16 (5)	20-16 (5)	non	occurrence				
100-91 (1)	15-11 (6)	non	occurrence					20-16 (5)	15-11 (6)	-3,7705	0	0	0	0	0
100-91 (1)	10-6 (7)	non	occurrence					20-16 (5)	10-6 (7)	-0,4504	0	0	0	-1,9625	0
100-91 (1)	5-0 (8)	non	occurrence					20-16 (5)	5-0 (8)	-2,1898	0	0	0	1,349	0
100-91 (1)	TD (9)	-2,541						20-16 (5)	TD (9)	-7,1937	0	0	0	0	0,07932
100-91 (1)	Punt (10)	-5,08575	0	0,05326	0,05191	0	0	20-16 (5)	Punt (10)	FREQUENCY : 0.004504505					
100-91 (1)	FG (11)	non	occurrence					20-16 (5)	FG (11)	3,00864	0	0	1,21576	0	-0,07351
100-91 (1)	Turnover (12	-2,3857	0	0	0	0	0	20-16 (5)	Turnover (12	-1,7346	0	0	-2,3597	0	0
90-61(2)	100-91 (1)	-16,3683	0	0	0	0	0,12907	15-11 (6)	100-91 (1)	non	occurrence				
90-61(2)	90-61 (2)	-0,69721	0	0	0	0	0	15-11 (6)	90-61 (2)	non	occurrence				
90-61(2)	60-41 (3)	-0,59816	0	0	0	0	0	15-11 (6)	60-41 (3)	non	occurrence				
90-61(2)	40-21 (4)	-3,02023	0	0	0	0	0	15-11 (6)	40-21 (4)	FREQUENCY : 0.003921569					
90-61(2)	20-16 (5)	-6,2687	0	0	0	0	0	15-11 (6)	20-16 (5)	FREQUENCY : 0.003921569					
90-61(2)	15-11 (6)	-5,3906	0	0	0	0	0	15-11 (6)	15-11 (6)	non	occurrence				
90-61(2)	10-6 (7)	-7,1861	0	0	0	0	0	15-11 (6)	10-6 (7)	-19,08593	0	0	0	0	0,2019
90-61(2)	5-0 (8)	-7,88	0	0	0	0	0	15-11 (6)	5-0 (8)	-2,240347	0,020868	0	0	0,808415	0
90-61(2)	TD (9)	-4,8868	0	0	0	1,3983	0	15-11 (6)	TD (9)	0,602272	-0,027879	0	-0,891295	0,680211	0
90-61(2)	Punt (10)	-2,882873	0	0	0	0	0,019953	15-11 (6)	Punt (10)	non	occurrence				
90-61(2)	FG (11)	non	occurrence					15-11 (6)	FG (11)	-1,3312	0	0	1,0484	0	0
90-61(2)	Turnover (12	-2,692921	0	-0,03734	0	0,40797	0	15-11 (6)	Turnover (12	-3,8738	0,0295	0,05687	0	0	0
60-41 (3)	100-91 (1)	non	occurrence					10-6 (7)	100-91 (1)	non	occurrence				
60-41 (3)	90-61 (2)	-5,332016	-0,027943	0	0	0	0,047513	10-6 (7)	90-61 (2)	non	occurrence				
60-41 (3)	60-41 (3)	-1,49306	0	0	0,21799	0	0	10-6 (7)	60-41 (3)	non	occurrence				
60-41 (3)	40-21 (4)	-0,700579	0,007593	0	0	-0,216872	0	10-6 (7)	40-21 (4)	non	occurrence				
60-41 (3)	20-16 (5)	-12,76866	0,0318	0,06216	1,75493	0	0,07491	10-6 (7)	20-16 (5)	FREQUENCY : 0.01052632					
60-41 (3)	15-11 (6)	1,9098	0	-0,08892	-0,80302	0	-0,06337	10-6 (7)	15-11 (6)	FREQUENCY : 0.01052632					
60-41 (3)	10-6 (7)	-3,9943	0	0	-1,2775	0	0	10-6 (7)	10-6 (7)	non	occurrence				
60-41 (3)	5-0 (8)	-4,1851	0	0	0	-1,8598	0	10-6 (7)	5-0 (8)	-24,27811	0,0472	0	4,20007	3,41915	0,20712
60-41 (3)	TD (9)	-2,98505	-0,02009	0	1,10985	0	0	10-6 (7)	TD (9)	5,36356	0	-0,09138	0	0	-0,06155
60-41 (3)	Punt (10)	-1,50447	0	0	0	0	0	10-6 (7)	Punt (10)	non	occurrence				
60-41 (3)	FG (11)	FREQUENCY : 0.003590664						10-6 (7)	FG (11)	-1,18462	0	0,08929	-0,67591	-2,1996	0
60-41 (3)	Turnover (12	-3,141	0		0	0,8539	0	10-6 (7)	Turnover (12	-3,9187	0	0	0	1,8392	0
40-21 (4)	100-91 (1)	non	occurrence					5-0 (8)	100-91 (1)	non	occurrence				
40-21 (4)	90-61 (2)	non	occurrence					5-0 (8)	90-61 (2)	non	occurrence				
40-21 (4)	60-41 (3)	-4,041	0	0	-1,137	0	0	5-0 (8)	60-41 (3)	non	occurrence				
40-21 (4)	40-21 (4)	-1,14286	0	0	-0,46883	0	0	5-0 (8)	40-21 (4)	non	occurrence				
40-21 (4)	20-16 (5)	-1,8219	0	0	0	0,3141	0	5-0 (8)	20-16 (5)	non	occurrence				
40-21 (4)	15-11 (6)	-2,889684	-0,17883	0	0	0	0,029021	5-0 (8)	15-11 (6)	FREQUENCY : 0.02314815					
40-21 (4)	10-6 (7)	1,87764	0	0	0,61713	0	-0,06738	5-0 (8)	10-6 (7)	non	occurrence				
40-21 (4)	5-0 (8)	-1,192962	-0,024219	0	-1,013398	-3,101984	0	5-0 (8)	5-0 (8)	6,37892	0	0	0	0	-0,12318
40-21 (4)	TD (9)	-4,033885	0,018112	0	0,912651	0	0	5-0 (8)	TD (9)	3,16475	-0,04347	-0,06633	0,78332	0	0
40-21 (4)	Punt (10)	-9,09184	-0,02153	0,04614	-0,65989	0	0,08961	5-0 (8)	Punt (10)	non	occurrence				
40-21 (4)	FG (11)	-2,486001	0,015649	0	0	-0,358972	0	5-0 (8)	FG (11)	-1,3633	0	0	0	0	0
40-21 (4)	Turnover (12	-2,3499	0	0	0	0	0	5-0 (8)	Turnover (12	-25,21846	0	0,18981	0	4,47014	0,2321

Table 3: Beta values

A total of 31 transitions are set to zero due to not occuring in the dataset. Furthermore, 7 transitions are determined using frequency analysis, either due to the testing algorithm or due to non-convergence of the *glm()* function in R.

The model based upon frequency analysis is denoted FA. Furthermore, 20 different covariate combinations were produced by the regression for different parts of the transition matrix. These models are denoted RM1 through RM20. The covariate combinations that they represent are shown in Table 4. An x in a cell denotes the presence of the corresponding covariate in that model.

	Intercept	Temperature	Wind	HFA	Q4	ODS
RM1	x					
RM2	x					x
RM3	x		x		x	
RM4	x	x				
RM5	x		x	x		
RM6	x	x				x
RM7	x				x	
RM8	x		x			
RM9	x	x			x	
RM10	x	x	x	x		x
RM11	x		x	x		x
RM12	x	x		x		
RM13	x			x		x
RM14	x	x		x	x	
RM15	x	x	x			
RM16	x	x		x	x	x
RM17	x		x			x
RM18	x		x	x	x	
RM19	x	x	x	x		
RM20	x		x		x	x

Table 4: Covariate combinations

5.2 The Transition Matrix

Each transition probability is computed using the coefficients specified in Table 3. By combining the information in Table 3 and Table 4, the regression model source for each transition probability in the transition matrix can be displayed in Table 5. Only the first 8 rows in the transition matrix are displayed, rows 9 through 12 represent the absorption states and are not dependent on the statistical analysis. A 0 in Table 5 indicates that the transition represented by that cell is non-occuring in the dataset.

	100-90	90-60	60-40	40-20	20-15	15-10	10-5	5-0	TD	Punt	FG	Turnover
100-90	RM1	RM4	RM3	RM2	0	0	0	0	RM1	RM5	0	RM1
90-60	RM2	RM1	RM1	RM1	RM1	RM1	RM1	RM1	RM7	RM2	0	RM3
60-40	0	RM6	RM8	RM9	RM10	RM11	RM8	RM7	RM12	RM1	FA	RM7
40-20	0	0	RM8	RM8	RM7	RM6	RM13	RM14	RM12	RM10	RM9	RM1
20-15	0	0	0	RM2	0	RM1	RM7	RM7	RM2	FA	RM13	RM8
15-10	0	0	0	FA	FA	0	RM2	RM9	RM14	0	RM8	RM15
10-5	0	0	0	0	FA	FA	0	RM16	RM17	0	RM18	RM7
5-0	0	0	0	0	0	FA	0	RM2	RM19	0	RM1	RM20

Table 5: Model source for each transition probability

Each of the rows are normalized to sum up to 1 so that they represent a probability distribution.

The actual values in the transition matrix depends on the covariate values. An example is shown in Table 6. In Table 6 the covariates have the following values:

- Temperature: 60°F
- Wind speed: 10 mph
- Home Field Advantage?: No.
- Fourth Quarter?: No.
- Opponent Defensive Strength: 70

	100-90	90-60	60-40	40-20	20-15	15-10	10-5	5-0	TD	Punt	FG	Turnover
100-90	0.0442	0.6992	0.0175	0.0189	0	0	0	0	0.0958	0.0137	0	0.1106
90-60	0.0007	0.3398	0.3626	0.0475	0.0019	0.0046	0.0008	0.0004	0.0077	0.1886	0	0.0455
60-40	0	0.0256	0.1910	0.4572	0.0070	0.0331	0.0188	0.0156	0.0155	0.1893	0	0.0455
40-20	0	0	0.0201	0.2820	0.1623	0	0.0644	0.0772	0.0582	0.0296	0.2047	0.1015
20-15	0	0	0	0.0082	0	0.0239	0.4130	0.1068	0.1722	0.0048	0.1120	0.1591
15-10	0	0	0	0.0042	0.0042	0	0.0076	0.2924	0.2752	0	0.2253	0.1911
10-5	0	0	0	0	0.0105	0.0105	0	0.0010	0.5329	0	0.4257	0.0194
5-0	0	0	0	0	0	0.0290	0	0.1203	0.5939	0	0.2556	0.0011
TD	0	0	0	0	0	0	0	0	1	0	0	0
Punt	0	0	0	0	0	0	0	0	0	1	0	0
FG	0	0	0	0	0	0	0	0	0	0	1	0
Turnover	0	0	0	0	0	0	0	0	0	0	0	1

Table 6: Example of transition matrix

The transition matrix in Table 6 is used in the Markovian model to produce the absorption probability distributions for each of the transient states. The results are displayed graphically in Figure 3.

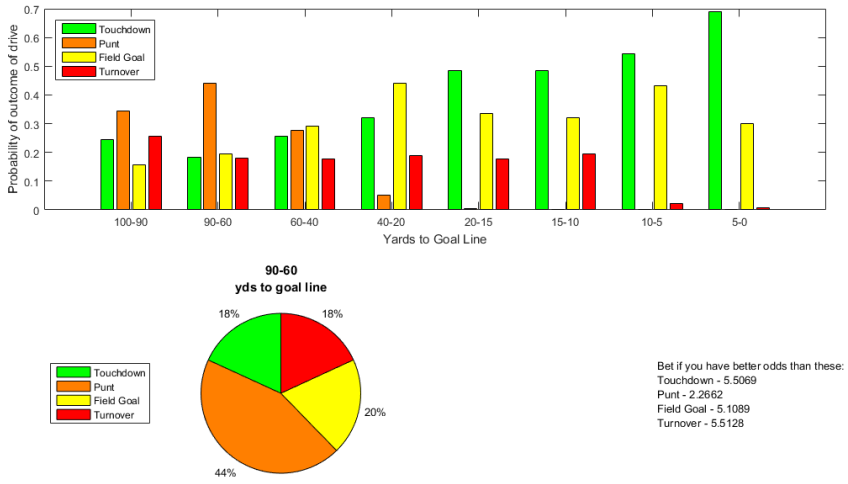

Figure 3: Results of calculations of absorption probabilities using the transition matrix in Table 6

Note that a set of odds is presented in Figure 3. These are the odds required for an expected profit to be made. E.g. if a first down is earned by the STEELERS in the 90-60 yard range and the bookmaker is offering odds of 2.5 for the event that the drive ends in a punt, then that bet should be made as $2.5 > 2.2662$.

Figure 3 represents a quite normal game scenario. The temperature and wind are moderate and the opponent's defense is of roughly average quality. It can be an interesting contrast to show a more extreme example. In Figure 4, the calculations are done with the following covariates:

- Temperature: 32°F
- Wind speed: 13 mph
- Home Field Advantage?: No.
- Fourth Quarter?: Yes.
- Opponent Defensive Strength: 100

This represents playing the league's strongest defense in their home stadium, in freezing weather and in the fourth quarter.

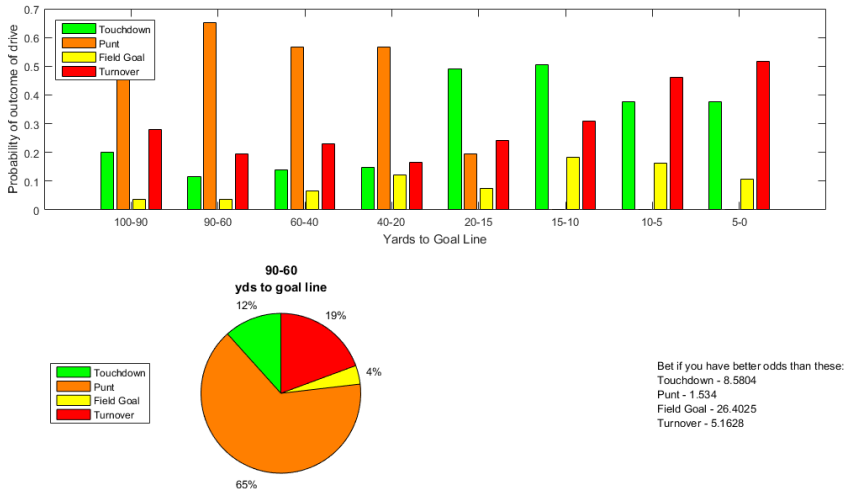

Figure 4: Cold weather, strong opponent

Notice the sharp contrast between Figure 3 and Figure 4. The probabilities of punting and turning the ball over are much higher in Figure 4, while the probability of scoring a field goal is understandably much lower. The odds required for profitable betting are of course also different. The same bet which required odds greater than 2.2662 in Figure 3 now only requires an odds of 1.534 to be profitable.

6 Analysis

6.1 Model Evaluation

The primary application of the Markovian model is producing statistical insight to help in live sports betting. Thus, the most obvious way of testing the model is to create a betting portfolio and make bets according to the advice of the model. If the portfolio grows then the model is outperforming the bookmakers in terms of predictive ability. Unfortunately, this thesis is being written during the spring of 2016. The NFL season runs every year from September to February. Thus, no games are played live during the spring which means that the model could not be tested in action during the process of writing this thesis.

The model can of course be tested against past seasons, there is no part in the model which requires the games to be live games. The problem isn't the lack of games but rather the unavailability of historical odds. There is no available database that stores live odds from entire games. Without the odds, no conclusions at all can be drawn regarding whether or not the model can be

profitable. The profitability is entirely dependent on the odds put forward by the bookmaker. In order to decide whether or not to make a bet, the bettor needs to know the probability of the bet being successful and the odds of the bet. The Markovian model can provide the probability, but the odds are impossible to know as they depend on the individual bookmaker's analysis of the probability.

6.1.1 Comparing the Regression and Frequency Analysis Based Models

In order to make some sort of evaluation however, one can consider the situation of comparing the performance of two predictive models. Note that such comparisons do not provide any information regarding the models' absolute profitability, but can give some insight into their relative predictive accuracy.

As mentioned earlier, the model can be used with a transition matrix based on frequency analysis instead of regression analysis. The matrix based on frequency analysis is much cruder and does not change depending on factors such as weather and opponent strength. The regression analysis matrix does so, however. As such, one would expect the regression analysis based model to perform better than the one based on frequency analysis. This can be tested using scoring rules.

The model is tested according to **4.3.4** with the 2015 season as testing environment. For every first down, a prediction is made by the regression based model and the frequency analysis based model. The predictions are scored using the logarithmic, quadratic and spherical scoring rules. The mean scores over the entire season are computed and displayed in Table 7.

	Regression analysis	Frequency analysis
Logarithmic	-1.2695	-1.3324
Quadratic	0.2999	0.2688
Spherical	0.5426	0.5200

Table 7: The mean scores for the regression and frequency analysis based predictions

Recall that the scoring rules are positively oriented, thus higher scores are better. According to Table 7 the model based on regression analysis scores higher irrespective of scoring rule. This means that the predictions made by that model were more accurate.

6.2 General Interpretation of Regression Results and Models

Robust and sound regression models to estimate transition probabilities can be regarded as the key to establishing a Markov model which approximates

an American Football game well. While this analysis won't include an in-depth structural interpretation of all the twenty regression models, a both quantitative and qualitative overview is put forth regarding each covariate and their estimates with regard to the hypotheses stated in **4.3.2**.

6.2.1 Temperature

The hypothesis states that temperature should affect the offense ability to pass the ball and force a more run oriented game resulting in fewer points and advancements. By examination of Table 3 in **5.1** it can be concluded that the covariate is significant only in 15 out of the 58 transition probabilities estimated by regression. The 15 regressions are distributed between 8 field position transitions and 7 transitions to an absorbing state.

From	To	Temp Estimate	From	To	Temp Estimate
100-91 (1)	90-61 (2)	0,0343	60-41 (3)	TD (9)	-0,02009
60-41 (3)	90-61 (2)	-0,027943	40-21 (4)	TD (9)	0,018112
60-41 (3)	40-21 (4)	0,007593	40-21 (4)	Punt (10)	-0,02153
60-41 (3)	20-16 (5)	0,0318	40-21 (4)	FG (11)	0,015649
40-21 (4)	15-11 (6)	-0,17883	15-11 (6)	Turnover (12	0,0295
40-21 (4)	5-0 (8)	-0,024219	15-11 (6)	TD (9)	-0,027879
15-11 (6)	5-0 (8)	0,020868	5-0 (8)	TD (9)	-0,04347
10-6 (7)	5-0 (8)	0,0472			

Table 8: Temperature coefficient estimates

According to the hypothesis, advancing field position estimates should be positive and loss of field position should be negative as warmer weather should favor the offense. Six out of the eight estimates are aligned with the hypothesis while two produce contradictory results. Both deviant estimates concern advancement from the 40-21 yard zone of the field close to the endzone. It could be that warmer weather and thus increased offensive abilities enables the offense to score touchdowns instead of being tackled close to the goal line. Hence, the warmer weather would reduce the propensity of advancing close to the goalline in favor of an increased probability for touchdowns.

A cursory glance of the temperature estimates gives some support to the aforementioned statement about the probability trade-off. The probability for a touchdown does indeed increase with the temperature, from the 40-21 yard zone. However, one must be careful to draw such a conclusion since the rest of the dataset that concerns touchdowns yields contradictory results. The contradiction is manifested by negative estimates. This implies that probability to score a touchdown decreases from certain spots on the field. It could be sheer happenstance that the touchdowns were few when at certain spots of the field when the weather was warm but the large amount of data points refutes that claim. The conclusion must be drawn that further research is warranted

to understand the contradictory estimates. Estimates for the punt and field goal transition does support the hypothesis. The transition to turnover seems to increase which again is odd if one holds the hypothesis to be true. This also warrants further research.

The overall low presence of temperature as a significant covariate in the regression models could lend credibility to the claim that temperature does not sufficiently explain events in a football game. It could be that there exist other underlying factors which may or may not be correlated to temperature that affects the game more, see **6.3**.

6.2.2 Wind

There are only 11 times out of 58 in which wind is found to be a significant factor. Three instances concerned advancement transitions and eight concerned absorption transitions. The hypothesis states that wind increases the difficulty of passing and kicking the football. Turnovers (specifically interceptions) increases as passing gets difficult. The first two transition estimates contradict the hypothesis. The probability of advancement increases with the wind. However, the third estimate is true to the hypothesis.

The examination of the transitions to absorbing states yields contradictory results as well, with respect to the hypothesis. Both punt and field goal probabilities increase with the wind. Whereas the punt probability is consistent with the hypothesis, the field goal probabilities should decrease with the wind as it is harder to kick in high wind speeds. Furthermore, it is interesting that wind is significant only when a field goal attempt is tried from such a short distance to the goal posts, out of all possible field goal distances. It is interesting because one could easily assume that wind affects less when the kick is shorter because it is easier to aim and the ball travels a shorter distance through the air. Hence, there is no easy logical explanation for this occurrence. Moreover, touchdown and turnover probabilities alter signs on their estimates which is not consistent with the hypothesis.

An explanation to this could be attributed to the ambiguity of wind. In the model, wind is assumed to be constant during the entire game which is obviously a false assumption. Wind can fluctuate plenty with wind gusts and short periods with no wind. Furthermore, it is implicitly assumed in the hypothesis that the wind described is a headwind or crosswind. Wind can obviously change direction during the course of a game and whereas headwind or crosswind could negatively impact passing and kicking, tailwind should have the opposite effect. For instance, the probability of kicking longer field goals should increase with a tailwind. It could simply be that the actual wind in the game differed in such a way that it impacted the transitions contrary to the hypothesis. As a covariate, further clarification in terms of information regarding the behavior of wind is needed to properly capture the effect wind should have upon a football game, see **6.3**.

6.2.3 Home Field Advantage

Nineteen regressions have home field advantage included as a significant factor that impacts the transition. The hypothesis regarding home field advantage suggests that the team that plays at home should perform better overall. All "negative" events such as punts and turnovers should decrease while all "positive" events such as touchdown, advancement and field goal should increase.

A initial review seems to reveal a contradiction towards the hypothesis with respect to advancement transitions. However, the negative estimates are for transitions from a considarable distance away from the endzone to zones close to the line of scrimmage. Such transitions are rare and indicates a good play by the offense. Home field advantage could push the offense to perform better in such a way that they score touchdowns instead of being tackled close to the goalline. As with temperature, it could be that there is a negative correlation between large advancements and touchdowns.

Touchdown estimates from corresponding zones on the field are consistent with the suggested explanation mentioned above. In the list there are a few outliers which requires further inspection. Punts should not be more likely with home field advantage from within the teams own 10 yardline. Furthermore, touchdown chances should not decrease while the team has the ball between the 20-16 yardline. The chance reduction of a field goal from within the 10 could be due to a trade off to an increase in touchdowns and/or advancement instead, i.e. the defense can't stop the offense as often and force a field goal when the offense has home field advantage. This is supported by the positive advancement estimate from the 10-6 yardline but since the covariate is not significant in the regression that concerns touchdown from the 10-6 yardline no further conclusions can be drawn.

6.2.4 Fourth Quarter

Nine advancement transition probabilities and eight transitions to absorbing states have the dummy Q4 included as a significant covariate. The hypothesis claims that a team in the fourth quarter should have more big plays, touchdowns, field goals and turnovers and less short plays and punts.

Again, the estimates produce mixed results with respect to the hypothesis. The first two estimates are consistent with respect to the trade off between large advancements as opposed to shorter ones. However, all other advancement estimates contradict the hypothesis. The decreased probability of advancing close to the endzone from far out could again be due to the fact that the offense might score touchdowns instead, because of its increased risk-taking. Interestingly, the probabilities of advancing from within the redzone to the endzone are all higher in the fourth quarter. An explanation could be that once the offense is within the redzone it reduces its risk taking and tries to advance more safely towards the endzone. Alternatively, it could be due to the defense being so concerned about stopping the offense to score that it adopts a strategy which protects the endzone but is less efficient at stopping shorter advancements.

Absorption probability estimates are consistent with the hypothesis except the field goal estimates. It's reasonable to believe however that if a offense is down by multiple scores it might be more inclined to try and go for it on fourth down and seek a touchdown.

A team in the fourth quarter should change its style of play whether it is in the lead or behind. However, it is not clear as exactly how the style of play changes as the estimates illustrate. It could be that a team in the lead does change its style of play more than anticipated with an increased propensity to run and punt the ball away. Although, none of the transitions to the punt state has Q4 as a significant variable. Subsequently, score differential in the fourth quarter might be a better covariate to anticipate the style of play, see **6.3**.

6.2.5 Opponent Defensive Strength

Intuitively, this covariate should have plenty of predictive power. A strong defense should severely impact the offense ability to score and advance down the field, regardless of where the offense is. Punts and turnovers should increase. Still, the defensive strength is significant only in 17 out of the 58 regressions. Eleven concerns advancement transitions and six are related to transitions to an absorbing state.

There are four estimates which are not conducive to the hypothesis regarding defensive strength. Two estimates are from field positions far from the endzone into the redzone and further investigation is required to determine the cause. The other two estimates are from within the redzone close to the endzone. These could be positive because a good defense rather surrenders yards than a touchdown. Subsequently, touchdowns from this these distances should decrease.

The touchdown estimate from the 10-6 yardline does indeed indicate that increased defensive strength decreases the probability of a touchdown from that distance. Inexplicably however, the touchdown estimate from the 20-16 yardline is positive. There is no intuitive reason for this and must be further researched. The punt, field goal and turnover estimates do support the aforementioned hypothesis regarding defensive strength.

It can be concluded that it remains unclear how the defensive strength of the opponent interacts with the offensive probabilities. A source of concern is that the defensive rank is based on an average performance over an entire season. Obviously, defensive performance should vary across a season, most likely from game to game.

To summarize it can be established that all of the covariates has associated estimates which require further research. Furthermore, one could have expected to see the covariates be significant more often in the regressions. This raises a concern that the covariates might not be optimal to predict transition probabilities within an American football game. Moreover, the complexity and dependency between the response variables elucidates the hardship in how to properly interpret results. All this points to the fact that improvement in the basic regression model is due and that there could exist other covariates which

better predict the outcome in a football game

6.3 Improvement Opportunities

6.3.1 Regression Improvements

The analysis presented above regarding the covariates elucidates the fact that there exist plenty of improvement or further research opportunities within the regression model. Although not all are listed below, a few suggestions for improvement are given. Improvements are selected on grounds such as intuition, data availability, and contribution to model complexity. The goal is to have a model which predicts a football game sufficiently well but still is intuitive, easy to manage, and easy to use.

1. First and foremost, the regression should be performed on all teams to verify that the issues of inconsistent covariate measurements do not pertain to factors specifically related to the Pittsburgh Steelers. If such is the case, then an in-depth analysis of the team is required to properly understand the causality.

2. Measures such as Goodness-of-Fit and Effect size should be incorporated to better gauge how much variance the model actually explains.

3. A set of dummy covariates which describes the type of weather should be added. Rain and Snow has a severe impact on how a game is played. Temperature is likely a proxy for weather conditions sometimes in the current regression. Hence, a separation should help produce more reasonable estimates.

4. The wind covariate should be removed. It is only significant in 11 of the 58 regressions and provided several confusing and unintuitive estimates. This might be solved by including wind directions but concerns regarding data availability and the arbitrariness of wind makes this unfeasible.

5. Player injuries should be included as a covariate. While hard to quantify, injuries have a substantial impact on team capabilities. In particular, injuries to key players such as the quarterback or a defensive star player should drastically impact transition probabilities. Concerns exist with respect to data availability but that is solved if the covariate is restricted to injuries of well known star players.

6. The defensive ranking system should be elaborated. Innately, defensive strength should have the biggest predictive impact. The covariate could be modified or new covariates added with data such as average yards per play to see whether a defense gives up bigger plays or smaller ones. Furthermore, it might be wise to divide each season into smaller periods to capture fluctuations in performance. A defense which is good at stopping the run might perform better during the winter months when weather conditions causes the offenses to pass less and run more for example.

7. Score differential should be included. This data is directly gathered through the *nflgame* API described in **4.2.2**. An interaction term with Q4 could be added to capture the impact that score differential and limited amount of time have. Consequently, Q4 could be removed as a stand alone dummy as its effect should be captured by the interaction term.

7 Conclusion

By the result and thesis analysis, it can be concluded that the Markovian model suggested in this thesis does achieve a higher predictive accuracy than a model strictly based on frequency analysis. It can also be concluded that factors such as temperature, wind, home field advantage, fourth quarter, and defensive strength are statistically significant and affect transition probabilities in some of the transitions. Furthermore, the conclusion can be drawn that the factors which impact transition probabilities are complex and require further research to better accentuate and explain certain factor behavior.

8 Industrial Management Application

8.1 Creating a Tech Start-up

The online betting market has a global market volume of 41 billion USD as of 2015 and has grown every year.[36] Part of that growth is driven by a popularity increase in live-betting.[6, 14] In-play betting is the key driver of the 69% growth in betting on soccer in the UK. Although live-betting for American Football has not yet caught on to the same extent as soccer, it is reasonable to believe that there exists a budding live-betting market there as well.

The mathematical model produced in this thesis could function as a foundation, on which upon a tech start-up could be built. The purpose of the start-up would be to exploit the nascent live-betting market for American Football and present a technical solution which with good predictive power can be used to set or evaluate live-betting odds. The product is envisioned as an online application where the mathematical model is used to tell the user whether live-odds given by a betting company is favorable or not. The premise is that the model supersedes the models of the betting institutions and takes advantage of presumptive flaws and thus enables the user to bet when odds are in the user's favor. A cursory search on the internet indicates that no comparable product exists at the present.

The purpose and aim of the Industrial Management section of this thesis is to present and discuss key areas in which the tech start-up has to be successful in order to develop into an established firm. Three key areas have been identified based on research of why start-ups fail. The areas are *Financing, Marketing/Branding* and *Monetizaton*. Research show that failure in any of these three areas significantly increase the mortality rate of the start-up.[21] An excellent financing strategy is crucial in order to secure funds to keep the

company growing. Marketing & Branding defines the target audience and helps the company focus its efforts. The monetarization policy is essential in exploring the possible revenue streams. These are all areas that can be adequately planned for in advance. The thesis will aim to answer research questions critical to success in each of these areas.

Financing

- How should the start-up obtain financing in order to avoid liquidity issues?
- How should the start-up minimize the cost of capital?
- Which type of capital should the start-up prefer?
- Which type of capital structure is optimal?

Marketing/Branding

- Who is the proposed consumer of the product that the start-up intends to develop?
- How should the start-up market its services?
- Which marketing channels should be used?

Monetization

- What should be the source of revenue for the start-up?
- How do different monetization policies impact the business?

8.2 Financing

Financing is the act of obtaining capital for business activites or the act of investing funds into a business or security. From an entrepreneurial perspective, the former definition is more applicable.

Capital, as pertained to monetary funds in a firm can be categorized as either *internal capital* or *external capital*. Internal capital can be defined as capital from owners, owner equity or capital generated from the business itself. External capital can be defined as any capital that originates from outside the business itself. Primarily, external capital can be divided into debt and equity. Debt capital is commonly structured as a loan on a fixed amount over a fixed time period with the debtor paying a cost of capital, interest rate, to the lender. Equity capital is normally structured such that the equity investor gains a fractional ownership and possibly access to control features of the business.

The relative proportion of debt and equity in a firm is referred to as the firm's *capital structure*.[2, p.479] According to the theories of perfect capital markets by Modigliani and Miller, and the *Law of One Price*, entrepreneurs should be indifferent with respect to the choice of a particular capital structure.[2, p.481]

This is because cash flows from an unlevered business activity are equal to cash flows from a levered business activity. Since leverage in the capital structure increases the risk associated with equity ownership in the firm, the expected return of the equity changes and thus keeps the cash flow equal as cash flows are discounted with the expected return.[2, p.482]

In a perfect capital market, it is assumed that [2, p.508-509]

- Investors and firms can trade securities at competitive market prices equal to the present value of their future cash flows
- There are no taxes, transaction costs, or issuance costs associated with the trading of securities
- A firm's financing decision does not change the cash flows generated by its investment, nor reveals any new information about the investments.

However, assumptions regarding perfect capital markets rarely hold. The presence of taxes enables a levered firm to raise more capital initially because interest rates are deducted before taxation. Thus, part of the firm's earnings is shielded from taxation and paid directly to debt investors. The resulting effect is that the business is able to increase the total amount available to all investors, both equity and debt holders, and hence be able to raise more capital initially.[2, p.509-510] Theory states that a firm should increase its leverage, or debt-to-equity ratio, such that the interest paid is equal to the firm's earnings before interest and taxes (EBIT).[2, p.529] As with any investment however, the issue with increasing debt levels is the risk associated with it. A higher debt level increases the risk of $default$, an inability to meet payment obligations which may cause a firm to go into *bankruptcy*. A bankruptcy process entails plenty of direct costs (e.g. legal fees) and indirect costs (e.g. loss of firm value, customers, suppliers) for both the firm and the investors.[2, p.543-547] Hence, both the value of the tax shield and the potential cost of bankruptcy must be contemplated when selecting a capital structure. In this decision, the product of the firm must be accounted for. A product with a volatile market will cause uncertain cash flows and increase the risk of bankruptcy, as opposed to a product with a stable market.[2, p.550] The complexity of these considerations causes the entrepreneur and the creditors to be cautious of accepting debt into the firm's capital structure.[20]

Apart from the issues with leverage and bankruptcy, firms and investors must also consider consider agency costs associated with choice of capital structure. Agency costs are costs which arise when there are conflicts of interest between firm stakeholders.[2, p.553] Start-up firms and start-up investors primarily have to deal with two types of problems which can incur agency costs, *moral hazard* and *adverse selection*.[20, 28] Moral hazard can be defined as the situation where one party of a contract takes more risks because the other party bears the cost of the risks. In a start-up financing situation, moral hazard can be recognized in the problem that the entrepreneur may act unobserved and misuse or misallocated external funding for personal benefit.[28] Adverse selection is a concept

that can be summarized as the skepticism a buyer feels when confronted with a sales proposal from a seller since there is an information assymetry regarding the value of the goods sold. The buyer is uninformed about the seller's reasons for the sale. Hence, when a seller has private information about the value of the good, buyers will discount the price they are willing to pay.[2, p.566] Adverse selection is prominent in a startup-setting because firms are often unlisted and not required to provide information.[40] Also, the start-up firm's assets are generally intangible and knowledge-based.[20] While an entrepreneur understands the quality of a proposal, investors might have some difficulty in comprehending or disagreeing about its value. Alternatively, the investor may be suspicious that the entrepreneur wants to capitalize before negative impact regarding the start-up emerges.[28]

The issues of moral hazard and adverse selection negatively impact the ability for a start-up firm to obatin external funds.[28] Furthermore, it makes external capital costlier than internal capital [20] because investors require higher returns on capital since start-ups are perceived as riskier due to the information assymetry.[5] External equity is generally hardest to obtain since it carries the highest cost for the entrepreneur. This is due to the fact that higher levels of external equity financing exacerberate moral hazard and in the case of firm bankruptcy, equity holders get paid last.[20] For an entrepreneur, high level of external equity dilutes the retained ownership and thus lowers incentives for the entrepreneur to run the firm properly.[2, p.559]

The increasing cost of capital from internal funds to external equity creates a natural pecking order when choosing how to finance investments. First, internal capital is used and debt issued only when the internal funds are exhausted. The firm issues equity only when all debt options are explored. [20, 5] Research suggests that this procedure is followed by the most nascent firms [20] and that it is closely tied to the size and age of the start-up.[40]

8.2.1 Capital Structure for the Thesis Start-up

With regard to the small size and the lack of tangible assets, it seems reasonable to initially finance the start-up with internal funds. Owner equity should be the first source of financing since the model is currently unfinished and unable to generate revenue. The equity aims to cover initial operating expenses and marketing campagins. As revenue is generated, it is reasonable to reinvest the funds into the start-up to further business development. The advantage of using internal funds is that the cost of capital is low. There is no risk premium involved due to the lack of information asymmetry since investors (the owners) are fully informed of the firm's operations and financial trends. However, one of the prime reasons why start-ups fail is due to lack of funds.[21] Hence, it is realistic to assume that at some point external capital is required to continue business development. This is consistent with the pecking order theory and the financial growth cycle theory which claims that different capital structures are optimal at different stages in the firm's development.[40]

The start-up should try to obtain capital by debt after internal funds are

exhausted, according to the pecking theory. Information asymmetry such as moral hazard and adverse selection are now concerns to creditors as they are not intimately involved within the business. Generally, the entrepreneur has to disclose financial records and reports of high quality in order to reduce information asymmetry to obtain funds from a creditor.[20] Furthermore, the creditor typically requires the entrepreneur to retain a large enough ownership stake in the firm to ensure that interests are alligned.[20] Hence, one can argue that a significant amount of time has to be devoted to secure debt funding. Time is not the only problem with debt in the capital structure of a start-up. Unlike equity, interest payments must be paid in a timely manner and paid regardless of whether the firm is flourishing or failing. Consequently, debt can lead to underinvestment where the entrepreneur is more concerned with paying off debt than developing the business.[2, p.555][20] Conversely, there is a significant risk of bankruptcy and its associated costs as the revenue stream of the start-up is expected to be volatile initially.[2, p.550] However, research indicates that there exists a correlation between debt and strong performance of a firm. This is due to the signal debt sends to investors, which states that the owners feel confident about the firm's ability to pay interest in the future because they are confident about the firm's prospect of future earnings.[20] Additionally, the benefits of the tax shield will help the start-up raise more funds as well. Despite the benefits of debt, one can argue that the firm has to be of a certain size and have a fairly stable revenue stream to be comfortable with accepting debt into the capital structure.

As an addition to the pecking theory, it can be argued that firms should only accept debt into their structure if they have solid financials and the bankruptcy and time costs are small. Furthermore, one can argue that firms which have exhausted their internal funding need to pursue capital through other means than debt funding. A different approach should be used by this start-up since the owners have limited funds which probably are not sufficient to ensure a stable business. An approach could be the use of *informal investors*. An informal investor is an individual, typically affiliated with the entrepreneur through social networks, such as friends, colleagues, neighbours, etc. They generally have less professionalism in extending financing.[28] Informal investors do not require the same monitoring and control abilities that institutional investors require. Instead, informal investors rely on the direct social tie to the investor to reduce the issue of moral hazard. The moral hazard is reduced because of social norms of obligation and fairness which are induced due to the entrepreneur's wish to protect its reputation within the social network.[28] Moreover, informal investors are morel likely to have private information regarding the firm due to the social tie and thus reduce the problem with adverse selection. Furthermore, informal investors are more likely to invest in firms which have a new product [28], which is ideal in this case since the start-up product is new. Due to the characteristics of the informal investor, it's reasonable to believe that funding received from this source consist of smaller amounts compared to funding from institutions such as banks or formal equity investors. Thus, it is argued that informal investors should work only as a bridge between internal funds and debt financing.

After funds from informal investors have been used and the business has solidified, debt should be introduced into the capital structure. There should be enough financial data at this point to reduce information asymmetry between the firm and the creditors. This enables the start-up to enjoy the benefits associated with debt. Short-term debt should be preferred to long-term debt as it is easier for creditors to estimate the probability of failure short-term than long-term, thus reducing cost of capital on short-term debts for the firm.[20, 5]

Based on the pecking order theory one can argue that external equity should be rejected altogether while debt funding is available. However, the theory ignores the strong empirical links that exist between the presence of external equity and start-up size, time to new firm founding, and general success.[20, 7] Often, a prospectus is prepared for potential equity investors about the short-termed and medium-termed prospects of the business. It reduces the information asymmetry between the firm and investor. The prospectus also helps the firm to review and refine the business idea and asses its probability of success, which could identify potential improvements.[20] In addition, the presence of equity investors could generate a lot of vital information capital to the firm. The firm benefits through consultation and support from the investor.[20] Hence, one can argue that there exist incentives for the start-up to quickly seek external equity as soon as possible.

It is important to acknowledge that the financing strategy outlined above probably is influenced by plenty of tangible and intangible factors. For instance, it's been assumed that all financing options outlined are readily available. This assumption might not hold in all cases. Informal investors might be hard to identify and they might not be able to provide enough funds to ensure a stable business. Consequently, the firm might take on debt sooner than optimal. Moreover, it could be that an external equity investor with the right information capital does not exist with respect to the proposed business model. In that case, debt financing could be used as the primary source of external founding.

To summarize, the thesis start-up capital structure should vary over time. The thesis start-up should first use internal funds such as owner equity and firm revenue until exhausted to keep cost of capital at a minimal level. Then, the thesis start-up should seek cheap capital from informal investors such as family and friends to give the firm time to establish a stable revenue stream. The time also helps the firm gather enough financial data to reduce the information asymmetry and thus be ready for debt funding. Short-term debt should be sought after rather than long-term debt to reduce cost of capital. Once debt funded, the thesis start-up should start seek external equity. With debt funding providing a secure source of capital, the thesis start-up should take time to find an investor with experience and information regarding areas vital to the start-up such as marketing, internal operations, mobile applications etc. Only once such an investor is found, should the thesis start-up accept external equity into its capital structure.

8.3 Marketing

The purpose of marketing is to influence potential consumers to be more inclined to purchase the product or service that a firm has to offer. Broadly, marketing could be broken down into two issues; Who should the firm direct its marketing towards? How should the firm market its products or services?

There is a need to categorize consumers to know who to market to. A categorization could be performed in a variety of ways such as by gender, by age, or by nationality etc. The general idea is to create customer segments to capture a certain consumer behavior which exist within the segment. The firm then wants to adapt a marketing method which caters to the customer behavior. A way to categorize consumers is to view the technology S-curve and group consumers by where they are likely to purchase a product or service with respect to where the technology is on the S-curve.[32, p.55-57]

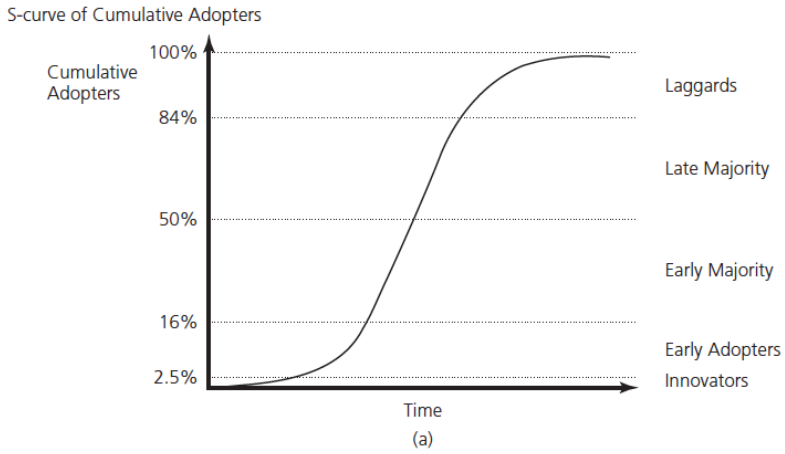

Figure 5: Technology S-curve

The figure indicates how a technology develops over time and where certain consumer groups are interested in the product. Each segment has certain characteristics:[32, 24]

- *Innovators* are characterized as adventurous and comfortable with a high degree of complexity and uncertainty. Introduce the technology into a social system but do not act as primary missionaries and opinionates for the technology.
- *Early Adopters* are highly integrated in the social system. They are well respected and know that they must make sound decisions to retain that respect. Are tremendous missionaries for new products or services due to the status they hold within the social system.

Most important for both Innovators and Early Adopters is the functional or technological advantage that the product or service provide over old technology. The financial advantage is less important as resources typically are not scarce.

- *Early Majority* accepts a product earlier than most. They look for products which are tested and will be the standard in the industry. Although not opinion leaders they do interact frequently with peers and thus are important to establish a solid customer base.
- *Late Majority* approaches new products and services with skepticism and may refrain from purchase until peer-pressured. Typically have scarce resources and as such are reluctant to invest in a good until all uncertainty is removed.
- *Laggards* is the last group to adopt a new technology. They are extremely apprehensive about new products or services and must feel ensured that a technology will not fail before purchase. Relies on past experiences rather than social groups and have no opinion leadership.

The most important factor for the Majority groups and the Laggards are the financial aspect and the supply security aspect. Products must be affordable and supplied by a credible or well-known supplier, which is a problem for unknown start-ups. The behavior shifts to a more pragmatic risk-evaluation view.[32, 24]

A conclusion drawn from the customer categories is that it is paramount for a firm to cater to different consumer segments at different points in the development of a technology. The characteristics of innovators and early adopters must be incorporated in marketing strategies as the business launches its product or service. As a product has launched and gained traction, the firm's marketing focus has to shift towards the early majority.[24]

8.3.1 Marketing Channels

When customers have been defined and strategies about how to take advantage of their characteristics are incorporated, questions remain about which marketing mediums a firm should use. An issue for start-ups is that they typically face additional barriers over established enterprises which affect the choice of marketing channel and advertisement formats. It stems from lack of financial and human resources, low corporate and brand awareness and risks of bankruptcy.[24] This must be considered when an appropriate marketing strategy is conceived.

To combat these issues, the paper *"Guidelines for e-Startup Promotion Strategy"* by D'Avino et al.[9] suggests a three step approach which filters potential advertisement formats down to the optimal ones.

1. At the first step, feasibility of each format is evaluated with focus on adaptability to market, e.g. a format which is too technologically advanced for a market should be avoided. The market in this model is divided along nations. The ICT development index (developed by the Information Telecommunication Society) is used as a measurement for each nation's

level of technology. This is compared to a IDIF value, which analogously ranks common advertisement formats with respect to technology level. All formats which has a IDIF value lower than a nation's ICT development index are deemed feasible in the specific country.

2. In the efficiency stage, formats which are too expensive are filtered away. The cost of the format conditional on the target market size is compared to the advertisement budget. The format is discarded if the cost is greater than the budget. The paper recognizes the difficulty of establishing the budget level but suggests that it could range from 5% to 30% of available firm resources.

3. Remaining formats are evaluated with respect to the cost / impact – ratio. Although it is difficult to evaluate impact, advertisement formats which offer the best combination of high impact and low cost should be chosen.

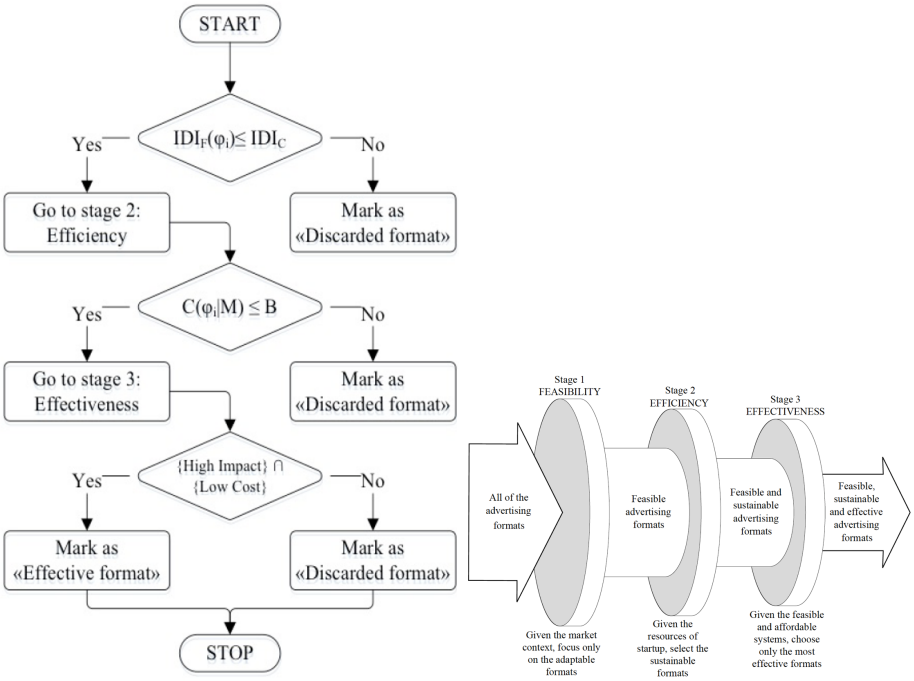

Figure 6: Three-step approach from *Guidelines for e-Startup Promotion Strategy*, D'Avino et al.

8.3.2 Marketing Strategy for the Thesis Start-up

An amendment to the marketing theories is required to properly implement a good marketing strategy. First and foremost, this strategy put forth below is

alligned towards end consumers and not betting institutions. While the latter are clearly viable consumers, it is reasonable to believe that creation of a firm as such is not needed as it is only a matter of selling the mathematical algorithm. One could envision the start-up to become a betting company of its own but it is deemed unfeasible due to immense competition and narrow product portfolio. Secondly, the strategy must also recognize in which countries the end consumers exist. This is required by the three step model advertisement evaluation model and to further the understanding of customer behavior.

The entire world could be viewed as a potential market due to the product being an online application. To reduce scope an initial selection is made. Tentatively, the United Kingdom could function as a good initial market.

- The United Kingdom ranks 10^{th} in gambling / capita which indicates a high interest in gambling.[1]
- Interest for American Football is high in the UK. The NFL plays three games in London each year in front of large crowds.[27]
- London is regarded as one of the top technology start-up cities in Europe with a good eco-system of potential investors and collaborators.[41]

The USA is not deemed viable because it outlaws gambling in a majority of its states and sports gambling is not regarded as high as legal casinos and lotteries.

Based on UK as the initial market, the best advertisement format for the start-up can now be selected through the three step model outlined above. The ICT development index for the UK is 8.50 as of 2013.[39, p.99]. Table 9 and the selection algorithm indicates that all advertisement formats are feasible for customers in the UK. The cost of the advertisement format and the budget constraint must be known to establish if a format is efficient or not. However, it is difficult to predict the budget constraint as the start-up does not yet exist. Still, since the owners are students one can argue that advertisement budget must be kept as low as possible. Hence, it is reasonable to select Blogging, Social Media, Digital Video, and Mobile Applications as efficient advertisement formats. The next step is to select the effective formats. While an in-depth analysis of each format is out of the thesis scope, intuitive arguments could be made to discard Blogging and Digital Video formats. Blogging does not seem to be an appropriate medium towards gamblers as it would be more towards customers interested in fashion and gadgets. Hence, impact could be fairly low. Digital Video on the other hand could have some impact but requires special video production competence which could be costly to acquire. Thus, cost could be fairly high. Therefore, the preferred advertisement formats should be Social Media and Mobile Applications. They arguably offer low costs with good customizability because advertisement usually can be directed towards certain users, thus increasing impact.

Ad Format	Marketing Channel	Instrument Required	Estimated IDI_z Range
Outdoor	Traditional	Nothing	0
Newspaper		Paper	1 / 2
Magazines			
Radio		Radio	2 / 3
TV		TV	3 / 4
Mobile (SMS)		Mobile phone	4 / 5
Sponsorship	Web 1.0	Computer + Internet (with slow connection)	
Display Ads			5 / 6
Rich Media			
Direct Mail			6 / 7
SEO & SEM			
Blogging	Web 2.0	Computer + Internet (with fast connection)	7 / 8
Social Media			
Digital Video			
Mobile (App)		Smartphone	8 / 9

Table 9: Advertisement formats

The formats should then be the primary way for the start-up to influence customers. The customizability makes them applicable throughout the product life-cycle. At launch, the product will be technology driven and less driven by market demands due to lack of comparable odds evaluator products which could have provided insight into customer demands. This could turn away customers which are characterized as early majority to laggards as they prefer refined and market driven products. Thus, key is to initially target the innovators and the early adopters. Their main interest could be to test whether the product actually predicts American Football games well rather than making a profit of it. The advertisement formats could be customized to target users of mobile gambling apps and gambling communities on social media sites such as Facebook, Twitter etc. Delivered content should focus on the technological superiority that the product provides. At this stage, less concern for the start-up should be of finding a credible supplier and instead focus on creating social media content which is interactive in order to gather feedback. The feedback is required to understand the market and create a market driven product which can bridge the chasm between the early adopters and the early majority.[24]

Assuming the product is a success with the innovators and early adopters and that proper feedback is received, marketing efforts should shift towards the early majority segment and the rest of the consumer segments. They ought to be more concerned whether profit can actually be made of the model as

financial resources are more scarce. Subsequently, the content advertised must now focus on cost advantages of the product.[24] Furthermore, marketing could be performed through a trusted supplier by the customers to reduce skepticism surrounding the product.

The actions outlined herein could function as a foundation to a market strategy. It is reasonable to conclude that further research and development is required, however. Advertisement formats must be selected using actual budget constraints and further developed. Furthermore, advantages and disadvantages of being first to market should be analyzed. Lastly, further research should investigate how marketing affects the ability to obtain financing. Several sources suggest that there exists a strong link between advertisement efforts and the ability to obtain financing.[24, 9]

8.4 Monetizaton

In order to create a long-term economically sustainable business model there must exist some source of revenue streams. The product must be able to generate a profit for the company and its owners. The choice of monetization policy is not so obvious however. There are many ways in which an application or website service can be monetized, each with its own strengths and weaknesses. In this section, a number of such monetization policies are explored and examined.

8.4.1 Advertisement

Free-to-use mobile applications and websites commonly generate revenue through advertisment of third-party companies displayed to the user. The obvious benefit of this is that the user avoids paying any money to use the application and website, but money is still generated to the developer. The user instead pays for the service by tolerating the presence of advertisement. Advertisements are provided by a third-party *ad network*. A company that wishes to advertise through applications and websites contatcs the ad network and pays them to market their products or services. The ad network then distributes the company's advertisements to its partners and the advertisement is displayed to some end-user. The ad network may categorize its advertisement outlets so as to be able to provide more targeted advetisement opportunities. For instance a company that produces fishing supplies may want to specifically have its advertisements displayed on fishing related websites and applications.

The effectiveness of advertisement in this context is measured using *effective cost-per-mille* (eCPM). For the owner of the website or application, eCPM meaures the advertising revenue generated per $1,000$ *impressions*. An impression is when an ad gets displayed on the website or application to someone in the world. eCPM is calculated as $\text{eCPM}=\frac{\text{Total Earnings}}{\text{Total Impressions}} \times 1000$.[31, p.66-67] Thus the total earnings generated by advertising depends on the eCPM and number of impressions. As such, there are two ways to increase the ad revenue:

- *Increase eCPM.* The eCPM depends on the what the ad network is willing to pay for the impressions. Essentially this can be seen as the price of

impressions. It may be difficult for a small start-up to have the negotiation leverage necessary to broker an increased eCPM though. The start-up would have to present compelling growth projections with promises of long-term partnerships, or perhaps be able to provide a very niche target audience, in order to be successful in such negotiations.

- *Increase number of impressions.* Far easier for the developer to micromanage is the amount of impressions generated by the website or application. Increasing the number of advertisement spots, and the rate at which they are displayed, will generally increase the number of impressions. However, this does not mean that cluttering the website or application with as many advertisements as possible is a wise decision. Too many advertisements harms the overall visual impression. If users decline to use the website or application due to there being too many advertisements, then an increase in advertisements will subsequently lead to a net decrease in impressions due to a loss of users. Thus the developer must find a balance between the number impression-generating advertisements, and having a clean user interface.
  Other than increasing the number of advertisements shown simultaneously, the developer can control the rate at which advertisements are displayed through the *refresh rate*. The refresh rate is how often a new advertisement is sent from the ad network for display.[31, p.68] A high refresh rate will generally result in more impressions, but a too extreme refresh rate will again be distracting for the user and could negatively impact the number of impressions.

The amount of advertisements shown can sometimes be limited by the ad network not having any client advertisements to distribute. The percentage of time that the application or website has ads when it is ready to show them is known as the *fill rate*. Naturally a high fill rate is desirable. A fill rate of 100% means that the ad network has an available advertisement whenever a request is made. Quite often, a high fill rate is a tradeoff against the eCPM. An ad network with a 100% fill rate operates at high volumes, meaning that they likely accept a wide variety of advertisements and advertising outlets. As such the eCPM can be expected to be lower. On the other hand, an ad network with a high eCPM could be one that specializes in a certain market segment. There the availability of advertisements is likely to be more volatile, leading to a fill rate of less than 100%.[26]

As an example, let's consider the monetization for a mobile application with the Markovian model presented in this thesis. The application is monetized using advertisements with the following parameters:

- Number of unique monthly users: $10,000$
- Monthly mean time of usage per user: 6 hours
- eCPM: \$0.80

- Fill rate: 95%
- Refresh rate: 1 advertisement request per minute

With a mean time of usage of 6 hours per user and month, the application will request $1 \times 6 \times 60 = 360$ advertisements per user and month. With $10,000$ users this equates to $3,600,000$ advertisement requests per month. If the ad network has a fill rate of 95%, this means that the application will display $0.95 \times 3,600,000 = 3,420,000$ advertisements per month, i.e. $3,420,000$ impressions per month. An eCPM of \$0.80 means that the ad network pays \$0.80 for every $1,000$ impressions. Thus the application will generate a monthly revenue of $\$0,80 \times 3,420,000 \div 1,000 = \$2,736$.

Pros

- No monetary cost for users, makes market entrance easier
- Easy to implement
- Requires little to no upkeep

Cons

- Requires a large amount of users to generate significant revenue
- Negatively impacts design of user interface
- Dependent on third party ad network

8.4.2 One-Time Charge

Instead of making the website or application free to use, the developer may choose to charge a one-time fee for buying the service. This is more common with applications than websites. The average price for a pay-to-use application on the *App Store* in January 2016 was \$1.16.[35] The user pays the fee in order to download the application and is then free to use it for as long as they like.

Two metrics decide the revenue generated by a pay-to-use application; price and number of downloads. Naturally, one can expect a correlation between the two due to supply and demand. Increasing the price is likely to decrease the number of downloads as less users are willing to pay a higher price. Finding the optimal application price would require an extensive market analysis.

A problem with only monetizing using a fee for downloading the application is that there is limited incentive to have the users use the application much once it is downloaded. There is no direct profitability benefit if the customer uses the application 24 hours a day over if the customer uninstalls the application directly after purchase. Of course, there are indirect benefits; a user who likes the application and uses it often is more likely to recommend it to other potential users. The incentive is weaker though, than in the case of advertisement monetization where the revenue is a direct function of time-of-usage.

Another major issue is that it may be difficult to enter the market if the application carries an immediate cost. The user has to pay upfront, prior to actually using the application. Therefore they must rely on external information about the quality of the application. Users are more likely to pay for the application if it has been recommended to them. However if the application is new to the market, the user is less likely to find reviews and recommendations and as such will have less information about the application prior to purchase. A remedy for this is to offer a free trial version of the application. The trial version gives the user access to some, but not all of the functions of the application. If the user likes the application then they may choose to pay for the full version. Here, the developer may choose to also include advertisement monetization. The trial version of the application could include advertisements. As the user isn't paying for the trial version, they will be more tolerant of advertisements than they would in the full version. It should also be communicated clearly to the user that paying for the full version will remove the advertisements. The issue with having a trial version of the application is that it may steal customers from the full version in the sense that some users will be content with only using the trial version and never actually paying for the full version. Thus, the trial version must be elaborate enough to give the user a taste of the full product, but not to the point where it removes the incentive for purchasing the full version.

Pros

- Easy to implement
- Requires little to no upkeep
- Can be combined with advertisement monetization in a trial version

Cons

- Difficult to enter market
- Users must rely on external information prior to purchase
- Dependent on a continous stream of new users
- Low incentive to have customers use the app for an extended amount of time

8.4.3 Subscription Fee

Another option for monetization would be to have users pay to use the betting service for a limited amount of time, for example by paying a monthly subscription fee. This ensures a steady stream of revenue over time that is less reliant on continuously attracting new users and more focussed on maintaining the current suscriber base.

A commonly employed strategy is to allow a free trial period of perhaps one month to let users test the website or application. If they like it then they continue using it and pay a monthly subscription fee. If they don't then they cancel the subscription after the free trial. By having the source of revenue come from existing users instead of new users, the developer is encouraged to focus on user experience instead of external product appeal. This promotes customer loyality which can be extended to other services provided by the same company. However, this also means that the developer is punished for neglecting user satisfaction and must therefore spend more time managing customer relations.

The total revenue generated by this monetization policy depends on the subscription fee and the number of subscribers. As with the application fee in **8.4.2,** a higher subsciption fee will make less people willing to subscribe. Therefore careful analysis must go into determining the price level. Unlike the case with one-time upfront fees, a subscription fee makes the future revenue easy to forecast and much less volatile. This is because the revenue depends on the amount of current users, not on the amount of expected new users.

Pros

- Steady and predictable stream of revenue
- Developer can focus on current subscriber base
- Promotes customer loyality

Cons

- Developer must spend time managing customer relations
- Must accumulate a sizeable subscriber base in order to generate significant revenue

8.4.4 Commission Fee

An interesting option for monetizing a betting advising service would be to charge a commission on the bets won by the user. The user pays nothing to download or be able to use the website or application. Instead, whenever they place a bet and win, they are charged a percentage of the net returns.

The intuitive benefits are appealing; if the user pays a commission fee when they win bets then the developer is highly encouraged to deliver as excellent a product as possible. A predictive model of higher quality will result in more won bets and thus a higher revenue. The revenue depends on the commission percentage, sum of bet stakes and percentage of total stakes that result in a won bet. This means that the revenue not only depends on how often the bets are won, but also on how large the bet stakes are. Therefore there exists an incentive to attract bettors that place large bets. However, larger bets mean larger commission fees if the commission is percentage based. It could be profitable

to offer a lower commission charge to bettors who bet with large stakes as the magnitude of their stakes will weigh up the lost commission percentage.

The fact that the commission only is charged when the bettor wins a bet is also appealing. The bettor should be more likely to accept having to pay a commission when they've won a bet as they will be making a net profit from the bet regardless. If the bettor loses a bet they still lose the money placed in stake, but at least they don't pay anything for the betting tool.

This monetization policy is much more difficult to technically implement though. In order to charge a commission, the bets must be placed through the website or application and not directly to the bookmaker. This requires some sort of technical integration into the bet placing system of the bookmaker. The bookmaker is unlikely to willingly let a betting tool be integrated into their betting site if the betting tool is profitable for the bettor. Thus the betting tool must be integrated into the betting process without the consent of the bookmaker. This presents a whole array of legal complications which lie beyond the scope of this thesis.

Pros

- Rewards technical excellence in the betting model
- Bettor is only charged when they win bets
- Cost for user scales with bet stake

Cons

- Technically and legally difficult to implement
- Cost for user scales with bet stake

8.4.5 Monetization for the Thesis Start-up

In order to decide the monetization policy most suitable for the thesis start-up a brief analysis of the target market is needed. The user is unlikely to have the application or website active when there is no current live game going on. If the user were to watch games during every time slot in a regular week they would watch three consecutive games on sunday and one game each on monday and thursday. Each game lasts roughly three hours, amounting to a total of 15 hours in a week and approximately 60 hours in a month. This represents the maximum reasonable usage time for one month, but the monthly mean time of usage is likely to be much lower. Thus it is safe to say that the average user would not have the application or website activated a very significant amount of time. For this reason, monetization by advertisement seems like a poor choice. Low time of usage will lead to a low amount of impressions and thus a low ad revenue.

The average user is likely to be quite cost aware. A bettor is looking to earn money by making sound betting decisions. Because the bookmaker already takes

a hefty cut, the bettor will be suspicious of any source of additional costs in their betting. As long as the increased profit generated by following the betting advice of the model outweighs the cost of using it, this won't be a problem. Assuming the model is able to give profitable betting advice, the problem instead is to convince new users to try it. For this reason, if the choice is made to have a subscription fee or one-time fee, there should be a free trial version available.

A subscription fee seems like a more suitable choice for monetization than a one-time fee in this case. The one-time fee is reliant on a steady stream of new users. However the idea of an American football betting tool is quite niche. It is not feasible to expect a large continous stream of new users over time. Because the target market is relatively small though, a subscription fee is an appropriate choice as the developer has a lot to gain from building healthy customer relations. The idea of a commission fee is very interesting, but seems too complicated to be immediately feasible. Thus it is recommended that the start-up have a subscription fee as the main source of revenue, however research into exploring the possibility of commission based monetization should be conducted.

8.5 Analysis

The suggestions presented in the three key areas above are deemed vital to establish a successful start-up, with regards to reasons why start-ups fail. Although vital, further analysis regarding the feasibility of the entire business is required. The analysis is performed by consideration of the technological potential in context with the aforementioned key areas. Lastly, certain development areas are identified to further increase chance of start-up success.

8.5.1 Feasibility of Thesis Start-up

A product which is in demand by the market is arguably the prime reason why a start-up is successful. Currently, the thesis start-up product is far from market and represents the biggest obstacle to overcome to establish the thesis start-up. Further development is required to create a product version with a user interface. Until such a development is made, the difficulty of marketing, obtaining financing, and generate revenue will be substantial as the product functionality is unable to be evaluated. Fortunately, plenty of intuitive app creation software is available on the market today with and as such this vital step can be viewed as feasible.

Strategies regarding marketing, financing and monetizaton can be implemented once initial technological progress is made. One can argue that there exists a co-dependency between the strategies with regard to feasibility. Capital is required o be able to start marketing the product through social media and mobile applications, hence there is a need quickly obtain financing. Conversely, to be able to obtain financing and external financing in particular, there is a need to have a developed market strategy in place. Advertisement efforts are not only noticed by users but by investors as well. The efforts could help persuade external investors to invest in the start-up, thus providing the necessary

financing to further develop the start-up. Although funds to advertise can be generated by the monetization method, it is dependent of a functioning marketing strategy in the first place. This complexity causes a Catch-22 situation which could question the overall feasibility of the business prospect. However, by a non-sequential implementation of each strategy in parts this complexity could be resolved. First, initial funding can easily be obtained from the owners and informal investors such as friends and family. This funding can in turn be used to market the product towards innovators and early adopters with a technology focused message through social media and mobile applications. A result should be a feedback loop from users of potential technological development. The further development should help accentuate the product making it more attractive to users and investors alike. Furthermore, the initial advertisement effort should provide some revenue through the subscription model with a free trial period might delay revenue stream. The steam can be used to further develop the product and to increase marketing budget. With this a feasible cyclical process is established which should lead to a full implementation of the strategies presented. If the strategies are implemented the start-up could expect to have a stable growth and a reasonable opportunity to establish as a proper firm.

8.5.2 Development Areas for Thesis Start-up

While strategy implementation within the three key areas should provide a solid foundation for the thesis start-up, there exist development areas. A cursory description is given as to what they are.

- *Internal Operations* – A need for additional employees might be required if the start-up grows to handle tasks of varying nature. Questions then arise regarding division of labor and responsibilities and how to manage and control these structures. Research should focus of finding an optimal firm structure for the thesis start-up to

- *Market Analysis* – There is a need to further accentuate who the prospective customers are. The model regarding customer segments along the S-curve should be applied, but there is also an important to regard other factors such that customers also are gamblers and interested in American Football. These characteristics could impact the advertisement efforts, e.g. a gambler might be more inclined to take on risk. An extensive market research should be conducted to get understand market size and to capture mentioned factors in consumer behavior.

- *Customer Feedback Loop* – An important feature of a start-up is to be responsive to its customers. Typically, it does not yet have enough momentum or a strong enough brand to ignore customers. Careful consideration must be placed on how to gather feedback from customers regarding the product and how to apply the information.

- *Revenue Stream* – The revenue stream is most likely to be highly cyclical as the Football season only runs from September to February. Consequently, the revenue should be high during this period and substantially decrease during the rest of the year. Further research should investigate whether this could pose a problem for the start-up or not.
- *Research & Development* – Arguably the central feature of a start-up. Continuous R&D is crucial to ensure that the product meet market demands. Emphasis must be put on establishing a creative environment where tools and support functions are tailored to further product development.

References

[1] Bartlett, W., Gbgc: Which country gambles the most? http://www.businesswire.com/news/home/20150610005010/en/GBGC-Country-Gambles, accessed: 2016-04-29.

[2] Berk, J. and DeMarzo, P. (2013) *Corporate Finance*. Pearson Education Inc., third edn.

[3] Bukiet, B., Harold, E., and Palacios, J. (1997) A markov chain approach to baseball. *Operations Research*, **45**, 14–23.

[4] Buse, A. (1982) The likelihood ratio, wald, and lagrange multiplier tests: An expository note. *The American Statistician*, **36**, 153–157.

[5] Cassar, G. (2004) The financing of business start-ups. *Journal of Business Venturing*, **19**, 261–283.

[6] Charlton, G. (2013), Uk's online gambling sector worth £2bn in 2012: Stats. https://econsultancy.com/blog/62407-uk-s-online-gambling-sector-worth-2bn-in-2012-stats/, accessed: 2016-04-28.

[7] Colombo, M. and Grilli, L. (2006) Start-up size: The role of external financing. *Economics Letters*, **90**, 148.

[8] Damour, G. and Lang, P. (2015) *Modelling Football as a Markov Process*. Master's thesis, KTH Royal Institute of Technology.

[9] D'Avino, M., De Simone, V., Iannucci, M., and Schiraldi, M. (2015) Guidelines for e-startup promotion strategy. *Journal of Technology Management & Innovation*, **10**.

[10] De Peuter, C. (2013) *Modeling Basketball Games as Alternating Renewal-Reward Processes and Predicting Match Outcomes*. Master's thesis, Duke University.

[11] Dubner, S. (2011), "football freakonomics": How advantageous is home-field advantage? and why? http://freakonomics.com/2011/12/18/football-freakonomics-how-advantageous-is-home-field-advantage-and-why/, accessed: 2016-04-26.

[12] Enger, J. and Grandell, J. (2003) Markovprocesser och köteori.

[13] Gabel, A. and Redner, S. (2012) Random walk picture of basketball scoring. *Journal of Quantitative Analysis in Sports*, **8**.

[14] Gainsbury, S. (2012) *Internet Gambling: Current Research Findings and Implications*. Springer.

[15] Gallant, A., nflgame. https://github.com/BurntSushi/nflgame, accessed: 2016-04-26.

[16] Gneiting, T. and Raftery, A. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, 359–378.

[17] Goldner, K. (2012) A markov model of football: Using stochastic processes to model a football drive. *Journal of Quantitative Analysis in Sports*, **8**.

[18] Goodell, R. (2015) *Official Playing Rules of the National Football League*. National Football League.

[19] Grinstead, C. and Snell, J. (1997) *Introduction to Probability*, chap. 11. American Mathematical Society, second edn.

[20] Hechavarría, D., Matthews, C., and Reynolds, P. (2015) Does start-up financing influence start-up speed? evidence from the panel study of entrepreneural dynamics. *Small Business Economics*, **46**, 137–167.

[21] Insights, C., The top 20 resons startups fail. https://www.cbinsights.com/research-reports/The-20-Reasons-Startups-Fail.pdf, accessed: 2016-04-28.

[22] Koski, T. (2014) Lecture notes: Probability and random processes at kth.

[23] Lang, H. (2015) Elements of regression analysis.

[24] Large, D., Grigorieva, E., and Falsetto, J. (2005) Best marmarket and sales practices for technology start-ups: a review and fresh evidence. *Proceedings. 2005 IEEE International Engineering Management Conference*, pp. 339–343.

[25] Liao, T. (1994) *Interpreting Probability Models: Logit, Probit, and Other Generalized Linear Models*. SAGE Publications.

[26] MonetizePros, Mobile ad cpm rates. http://monetizepros.com/cpm-rate-guide/mobile/, accessed: 2016-04-29.

[27] NFL, 2016 uk international series schedule announced. http://www.nfl.com/news/story/0ap3000000587776/article/2016-uk-international, accessed: 2016-04-29.

[28] Nofsinger, J. and Wang, W. (2011) Determinants of start-up firm external financing worldwide. *Journal of Banking & Finance*, **35**, 2282–2294.

[29] Nualart, D. (1997) Stochastic processes.

[30] Peña, J. A marmarkov model for association football possession and its outcomes.

[31] Rollins, M. and Sandberg, R. (2013) *The Business of Android Apps Development*. Apress, second edn.

[32] Schilling, M. (2013) *Strategic Management of Technological Innovation.* Richard D. Irwin, Inc., fourth edn.

[33] Shirley, K. A markov model for basketball.

[34] SportingCharts, Team plays per game: 2015 nfl season. http://www.sportingcharts.com/nfl/stats/team-plays-per-game/2015/, accessed: 2016-04-26.

[35] Statista, Average prices for apps in the apple apps store as of january 2016. http://www.statista.com/statistics/267346/average-apple-app-store-price-app/, accessed: 2016-04-29.

[36] Statista, Size of the online gambling market from 2009 to 2018 (in billion u.s. dollars). http://www.statista.com/statistics/270728/market-volume-of-online-gaming-worldwide/, accessed: 2016-04-28.

[37] Sykes, A. (1993) An introduction to regression analysis. *Coase-Sandor Institute for Law & Economics Working Paper*.

[38] Štrumbelj, E. and Vračar, P. (2012) Simulating a basketball match with a homogeneous markov model and forecasting the outcome. *International Journal of Forcasting*, **28**, 532–542.

[39] Union, I. T. (2014) Measuring the information society report.

[40] Vos, E., Yeh, A., Carter, S., and Tagg, S. (2007) The happy story of small business financing. *Journal of Banking & Finance*, **31**, 2648–2672.

[41] Wauters, R. (2015), London tops the list of most digital entrepreneur-friendly cities in europe. http://tech.eu/features/6439/european-digital-city-index/, accessed: 2016-04-29.

[42] Weber, R. (2011) Markov chains. Cambridge University.

[43] Willman, D., Nflsavant. http://nflsavant.com/index.php, accessed: 2016-04-26.

TRITA -MAT-K 2016:23
ISRN -KTH/MAT/K--16/23--SE